

КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

С.В. Симушкин

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

Часть II

Учебное пособие

Издательство
Казанского государственного университета
2009

УДК 517.9

Печатается по решению
Учебно-методической комиссии
факультета вычислительной математики и кибернетики КГУ

Научный редактор
доктор физико-математических наук, профессор И.Н. Володин

Симушкин С.В.

Многомерный статистический анализ. Часть II: Учебное пособие. — Казань: Казанский государственный университет, 2009. — 114 с.

В учебном пособии продолжается изложение курса многомерного статистического анализа, читаемого для студентов IV курса факультета ВМК, специализирующихся по кафедре математической статистики. Излагаются основные методы статистической обработки данных, содержащих большое число измеряемых показателей.

© Казанский государственный
университет, 2009

Оглавление

I	Статистические выводы о параметрах	4
§ 1.	Статистические оценки мер взаимосвязи	4
§ 2.	Асимптотические выводы	8
2.1	Критерии о коэффициенте корреляции	10
2.2	Преобразование Фишера	14
2.3	Наилучший линейный прогноз	17
2.4	Функциональная регрессия	22
2.5	Регрессия в многомерном случае	26
§ 3.	Точные распределения в нормальной модели	28
3.1	Моменты обобщенной дисперсии	29
3.2	Распределение коэффициентов корреляции	30
3.3	Преобразование Стьюдента	34
3.4	Проверка независимости групп признаков	36
3.5	Проверка независимости одного от всех	38
§ 4.	Доказательства	41
II	Ранговая корреляция	59
§ 1.	Коэффициент корреляции Спирмена	59
§ 2.	Коэффициент корреляции Кендэлла	62
§ 3.	Коэффициент конкордации	63
§ 4.	Доказательства	65
III	Дискриминантный анализ	68
§ 1.	Дисперсионный анализ многомерных данных	68
1.1	Проверка гипотезы о центре одной группы	70
1.2	Сравнение двух групп	71
1.3	Сравнение более двух групп	72
§ 2.	Классифицирующие правила	76
IV	Примеры	83

СТАТИСТИЧЕСКИЕ ВЫВОДЫ

О МНОГОМЕРНЫХ

ХАРАКТЕРИСТИКАХ СВЯЗИ

Г л а в а I.

В первой части курса [10] были введены числовые характеристики многомерных вероятностных распределений, позволяющие судить о степени взаимозависимости компонент случайного вектора. Статистический анализ начинается тогда, когда точное распределение неизвестно и выводы относительно того или иного параметра делаются по значениям ряда реализаций наблюдаемого случайного вектора. В этой главе мы найдем оценки интересующих нас характеристик и изучим их некоторые вероятностные свойства.

§ 1. Статистические оценки мер взаимосвязи

Большое количество информации
протицает мозги, но затуманивает
разум.

Предположим, что в каждом из n независимых экспериментов наблюдались реализации p -мерного случайного вектора $\vec{X} = (X_1, \dots, X_p)'$. Результаты наблюдений можно записать в виде матрицы

$$\mathbb{X}' = \begin{pmatrix} \vec{x}'_1 \\ \vec{x}'_2 \\ \vdots \\ \vec{x}'_n \end{pmatrix} = \begin{pmatrix} \overbrace{x_{11} \ x_{21} \ \dots \ x_{p1}}^{X_1 \ X_2 \ \dots \ X_p} \\ x_{12} \ x_{22} \ \dots \ x_{p2} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ x_{1n} \ x_{2n} \ \dots \ x_{pn} \end{pmatrix} \begin{matrix} \text{Джонсон} \\ \text{Петросян} \\ \vdots \\ \text{Сидорчук} \end{matrix}$$

Каждая i -ая строка $\vec{x}'_i$ этой матрицы представляет собой вектор значений наблюдаемых показателей в i -ом эксперименте (например, карта одного из

пациентов с совокупностью всех показателей здоровья), а j -ый столбец есть реализации j -ого показателя во всех экспериментах (например, содержание сахара в крови у всех пациентов).

Поскольку все рассматриваемые нами характеристики взаимосвязи (корреляция, регрессия, главные компоненты и т.д.) полностью определяются моментами первого и второго порядка, то оценки этих величин по методу моментов совпадают с их выборочными аналогами. Следовательно,

1) оценка вектора математических ожиданий $\vec{\mu}$ равна вектору выборочных средних $\vec{x}_{\bullet} = (x_{1\bullet}, \dots, x_{p\bullet})'$, где

$$x_{j\bullet} = \frac{1}{n} \sum_{i=1}^n x_{ji} \quad -$$

среднее j -го показателя (столбца), $j = \overline{1, p}$;

2) оценка матрицы ковариаций Λ равна выборочной ковариации

$$L = (l_{jk}) = \frac{1}{n} \sum_{i=1}^n \vec{x}_i \vec{x}_i' - \vec{x}_{\bullet} \vec{x}_{\bullet}' = \left(\frac{1}{n} \sum_{i=1}^n x_{ji} x_{ki} - x_{j\bullet} x_{k\bullet} \right)_{j,k=1}^p.$$

Легко видеть, что матрица L связана с матрицей исходных данных $\mathbb{X}$ и вектором средних соотношением

$$L = \frac{1}{n} \mathbb{X} \mathbb{X}' - \vec{x}_{\bullet} \vec{x}_{\bullet}'.$$

Для построения оценок по методу моментов любого из интересующих нас параметров, достаточно в формуле, определяющей этот параметр, заменить точные значения $\vec{\mu}$ и Λ их соответствующими выборочными аналогами. Например, оценки дисперсий равны

$$\hat{\sigma}_j^2 = s_j^2 = l_{jj} = \frac{1}{n} \sum_{i=1}^n x_{ji}^2 - x_{j\bullet}^2,$$

а оценки полных коэффициентов корреляции равны

$$\hat{\rho}_{jk} = r_{jk} = \frac{l_{jk}}{s_j s_k}.$$

З 1. Полный коэффициент корреляции между двумя рядами данных принято называть коэффициентом корреляции Пирсона.

При известном распределении наблюдаемого вектора признаков можно найти оценки максимального правдоподобия для изучаемых характеристик. В частности, для нормальной модели распределения вектора $\vec{X}$ справедлива

Теорема I.1. Если вектор наблюдаемых признаков имеет нормальное распределение $\mathcal{N}_p(\vec{\mu} | \Lambda)$, то

- (i) оценки максимального правдоподобия параметров модели совпадают с оценками по методу моментов;
- (ii) максимум функции правдоподобия равен

$$\max_{\vec{\mu}, \Lambda} f_n(\vec{x}^{(n)} | \vec{\mu}, \Lambda) = (2\pi e)^{-\frac{pn}{2}} |L|^{-\frac{n}{2}}. \quad (\text{I.1})$$

Теоретические моменты выборочных моментов.

Найдем выражения для математического ожидания и матрицы ковариаций выборочных вторых моментов. Для простоты восприятия формулировку утверждений будем проводить только для случайных величин X_1, X_2 .

Лемма I.1. Средние значения выборочных вторых моментов (l_{11}, l_{22}, l_{12}) равны

$$\mathbf{E}(l_{11}, l_{22}, l_{12}) = \frac{n-1}{n} (\lambda_{11}, \lambda_{22}, \lambda_{12}).$$

§ 2. Как видно из утверждения леммы, выборочные вторые моменты в среднем смещены относительно оцениваемых параметров. Очень часто на практике рассматривают несмещенные (исправленные) варианты этих оценок, получаемые из исходных путем умножения на $\frac{n}{n-1}$. Дисперсии и ковариации исправленных оценок, полученные в следующей лемме, уже

не будут содержать сомножитель $\left(\frac{n-1}{n}\right)^2$. При изучении асимптотических свойств (когда $n \rightarrow \infty$) оба этих сомножителя стремятся к 1 и, следовательно, не играют существенной роли.

Для дальнейшего нам понадобятся центральные моменты случайных величин X_1, X_2 . Обозначим через

$$a_{ml} = \mathbf{E}(X_1 - \mu_1)^m (X_2 - \mu_2)^l$$

$(m + l)$ -ый смешанный центральный момент. Тогда

$$\begin{aligned} \text{ковариация} \quad & \lambda_{12} = a_{11}, \\ \text{дисперсии} \quad & \sigma_1^2 = \lambda_{11} = a_{20}, \quad \sigma_2^2 = \lambda_{22} = a_{02}. \end{aligned}$$

§ 3. Индексы параметров λ_{kj}, σ_j^2 указывают на компоненты вектора $\vec{X}$, участвующие в их формировании, а индексы параметров a_{ml} указывают на степени компонент X_1 и X_2 .

Определим две симметричные матрицы размерности $3 \cdot 3$

$$\Omega = \begin{pmatrix} a_{40} & a_{22} & a_{31} \\ a_{22} & a_{04} & a_{13} \\ a_{31} & a_{13} & a_{22} \end{pmatrix}, \quad \Psi = \begin{pmatrix} a_{20}^2 & a_{20}a_{02} & a_{11}a_{20} \\ a_{20}a_{02} & a_{02}^2 & a_{11}a_{02} \\ a_{11}a_{20} & a_{11}a_{02} & a_{11}^2 \end{pmatrix}.$$

Отметим, что разность этих матриц

$$Q = \Omega - \Psi$$

есть ковариационная матрица вектора

$$((X_1 - \mu_1)^2, (X_2 - \mu_2)^2, (X_1 - \mu_1)(X_2 - \mu_2)).$$

Л е м м а I.2. Матрица ковариаций (l_{11}, l_{22}, l_{12}) равна

$$\mathbf{Cov}(l_{11}, l_{22}, l_{12}) = \frac{1}{n} \left(\frac{n-1}{n}\right)^2 \left(Q + \frac{2}{n-1}\Psi\right).$$

§ 4. Если пренебречь слагаемыми высшего порядка малости (при $n \rightarrow \infty$), можно сказать, что асимптотическая ковариация вектора (l_{11}, l_{22}, l_{12})

приближенно равна $\frac{1}{n}Q$.

З 5. Для нормальной вероятностной модели моменты связаны соотношением $a_{40} = 3a_{20}^2 = 3\sigma_1^4$ (см. лемму I.3, стр.14). Поэтому заявленная в предыдущей лемме дисперсия

$$\mathbf{D} l_{11} = \frac{1}{n} \left(\frac{n-1}{n} \right)^2 \left(a_{40} - a_{20}^2 + \frac{2}{n-1} a_{20}^2 \right) = 2(n-1) \left(\frac{\sigma_1^2}{n} \right)^2.$$

С другой стороны, из общего курса статистики хорошо известно, что для нормальной модели выборочная дисперсия l_{11} ($= s_1^2$) представляет собой хи-квадрат случайную величину, умноженную на $\frac{\sigma_1^2}{n}$: $l_{11} \sim \frac{\sigma_1^2}{n} \chi_{n-1}^2$. Если теперь вспомнить, что дисперсия $\mathbf{D} \chi_{n-1}^2 = 2(n-1)$, то получим совпадение результатов леммы с „прописными истинами“.

§ 2. Статистические выводы, основанные на асимптотических свойствах оценок

Счастье не в достижении цели,
а в самом движении к ней.
Сизиф

Построение статистических выводов невозможно без установления вида распределения выборочных статистик. Если не предполагается известной структура вероятностной модели наблюдений, то найти точное распределение, естественно, нельзя. Поэтому единственно возможным представляется путь, основанный на асимптотических свойствах оценок.

Закон больших чисел применительно к нуждам статистики можно переформулировать следующим образом.

Теорема I.2. Оценки вектора средних $\vec{\mu}^* = \vec{x}$. и матрицы ковариаций $\Lambda^* = L$ состоятельны.

§ 6. Поскольку все рассматриваемые нами многомерные характеристики представляют собой непрерывные функции от первых и вторых моментов, то их выборочные аналоги также *состоятельны*.

Основным инструментом для получения более глубоких свойств оценок служит так называемый „дельта-метод“, доказательство которого дано в первой части курса [10, стр.93, Теорема А3]. Ниже приведен усиленный вариант этого метода.

Теорема I.3. (Дельта-метод) Пусть

- k -мерный вектор $\vec{U}_n \rightsquigarrow \mathcal{N}_k(\vec{b}, \frac{1}{n}Q)$, $n \rightarrow \infty$;
- функция $h(\vec{u})$ дифференцируема в окрестности точки $\vec{b}$;
- $\vec{\phi}$ – вектор первых производных $h(\vec{u})$ в точке $\vec{b}$.

Тогда случайная величина $h(\vec{U}_n)$ асимптотически нормальна со средним $h(\vec{b})$ и дисперсией $\frac{1}{n} \tau^2 = \frac{1}{n} \vec{\phi}' Q \vec{\phi}$:

$$\sqrt{n} \frac{h(\vec{U}_n) - h(\vec{b})}{\tau} \rightsquigarrow \mathcal{N}_1(0, 1), \quad n \rightarrow \infty.$$

Справедлив многомерный аналог этой теоремы, когда функция h принимает значения в m -мерном пространстве. В частности, отсюда и из центральной предельной теоремы будет следовать, что вектор всех выборочных первых и вторых моментов асимптотически нормален.

Теорема I.4. Пусть $\vec{\lambda} = (\lambda_{11}, \lambda_{22}, \lambda_{12})'$ – вектор истинных вторых моментов, а $\vec{l} = (l_{11}, l_{22}, l_{12})'$ – вектор их выборочных аналогов, тогда при $n \rightarrow \infty$

$$\sqrt{n}(\vec{l} - \vec{\lambda}) \rightsquigarrow \mathcal{N}_3(\vec{0}, Q).$$

Комбинируя утверждение этой теоремы с дельта-методом, можно получить асимптотическое распределение любой функции от выборочных моментов. Например, для коэффициента корреляции Пирсона справедлива

Теорема 1.5. Пусть r_n – выборочный коэффициент корреляции Пирсона, построенный по выборке объема n с истинным коэффициентом корреляции ρ . Тогда при $n \rightarrow \infty$

$$\sqrt{n}(r_n - \rho) \rightsquigarrow \mathcal{N}_1(0, \tau_\rho^2),$$

где асимптотическая дисперсия равна

$$\begin{aligned} \tau_\rho^2 = & \frac{a_{22}}{\sigma_1^2 \sigma_2^2} + \\ & + \frac{\rho^2}{4} \left(\frac{a_{40}}{\sigma_1^4} + \frac{a_{04}}{\sigma_2^4} + 2 \frac{a_{22}}{\sigma_1^2 \sigma_2^2} \right) - \rho \left(\frac{a_{31}}{\sigma_1^3 \sigma_2} + \frac{a_{13}}{\sigma_1 \sigma_2^3} \right). \end{aligned} \quad (1.2)$$

З 7. Если параметры a_{jk}, ρ и σ_j^2 заменить их состоятельными оценками (например, выборочными аналогами), то оценка $\hat{\tau}_\rho^2$ дисперсии (1.2) будет состоятельной и в силу теоремы Слущкого функция $\sqrt{n}(r_n - \rho)/\hat{\tau}_\rho$ также будет асимптотически нормальна:

$$\sqrt{n} \frac{r_n - \rho}{\hat{\tau}_\rho} \rightsquigarrow \mathcal{N}_1(0, 1).$$

2.1 Асимптотические критерии проверки гипотез о парном коэффициенте корреляции

Проиллюстрируем способ применения предыдущей теоремы на примере проверки гипотезы $\mathbf{H} : \rho \leq \rho_0$ при альтернативе $\mathbf{K} : \rho > \rho_0$ о парном коэффициенте корреляции.

Для заданного уровня α определим $\mathcal{N}^{(\alpha)}$ как верхнюю α -квантиль стандартного нормального распределения: $\mathcal{N}^{(\alpha)} = \Phi^{-1}(1 - \alpha)$.

Теорема 1.6. Если r_n – выборочный парный коэффициент корреляции, $\hat{\tau}_\rho^2$ – состоятельная оценка дисперсии (1.2) при $\rho = \rho_0$, то при проверке гипотезы $\mathbf{H} : \rho \leq \rho_0$ против альтернативы $\mathbf{K} : \rho > \rho_0$ критическая область

$$\sqrt{n} \frac{r_n - \rho_0}{\hat{\tau}_\rho} > \mathcal{N}^{(\alpha)}$$

имеет асимптотический уровень α : $\forall \rho \leq \rho_0$

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \sqrt{n} \frac{r_n - \rho_0}{\hat{\tau}_\rho} > \mathcal{N}^{(\alpha)} \mid \rho \right\} \leq \alpha.$$

§ 8. На практике вместо вычисления квантили и сравнения с ней значения тестовой статистики гораздо удобнее вычислять критический уровень значимости критерия $\alpha_{\text{кр}}$ и сравнивать его с заданным уровнем значимости α . Чтобы определить $\alpha_{\text{кр}}$, заметим сначала, что любой критерий представляет собой, по-существу, семейство критериев, зависящее от критической константы, определяемой, в свою очередь, через заданное исследователем ограничение на вероятность ошибки 1-го рода – уровень значимости α . При этом, если для полученных экспериментальных данных критерий отвергнет проверяемую гипотезу при некотором выбранном значении α , то эта же гипотеза будет отвергнута и при больших значениях α (то есть при менее жестких ограничениях на вероятность ошибки 1-го рода).

Критический уровень значимости $\alpha_{\text{кр}}$ равен минимальному уровню значимости, при котором гипотеза отвергается для полученного в эксперименте значения тестовой статистики. Иногда обозначается буквой „ p “ и называется p -значением. Гипотеза должна быть отвергнута, если для выбранного ограничения α на вероятность ошибки 1-го рода (выбранного уровня значимости) выполняется неравенство

$$\alpha_{\text{кр}} < \alpha.$$

Обычно отвержение гипотезы означает принятие именно того утверждения, которое удовлетворяет исследователя. Поэтому малые значения

$\alpha_{кр}$ в большей степени свидетельствуют в пользу „ожидаемого“ утверждения. Исходя из этого, величину $(1 - \alpha_{кр}) \cdot 100\%$ часто интерпретируют как меру близости экспериментальных данных к альтернативе, а величину $\alpha_{кр} \cdot 100\%$ – как меру близости к гипотезе.

На практике часто не уточняется величина применяемого уровня значимости, а сообщается значение $\alpha_{кр}$, с которым гипотеза была отвергнута или принята. Например, если требуется установить наличие связи между дозой применяемого лекарства и артериальным давлением, то

- а) выдвигается противоположная гипотеза об отсутствии связи;
- б) проводятся статистические испытания;
- в) по результатам испытаний в соответствии с выбранным критерием вычисляется критический уровень значимости p ;
- г) если p мало (например, $p = 0.045$ или $p = 0.00000032$), то гипотеза отвергается, при этом делается вывод о том, что

наблюдается статистически значимая связь
с $p = 0.045$ (или $p < 0.001$)
(с надежностью 96.5% (>99.9%));

если p не достаточно мало (например, $p = 0.072$), то гипотеза не отвергается и делается вывод о том, что

связь статистически незначима
с $p = 0.072$
(с надежностью 7.2%).

Для нахождения критического уровня значимости критерия, основанного на тестовой статистике T , необходимо:

- i) записать вид критической области через критическую константу (например, $T > C$, $T < C$ или $|T| > C$);
- ii) по экспериментальным данным найти значение статистики $T = t_{эксп}$;
- iii) вычислить критический уровень значимости, заменив константу C на $t_{эксп}$, например,

$$\alpha_{кр} = \mathbf{P}\{T > t_{эксп}\}.$$

Строго говоря, последняя вероятность должна быть вычислена при

всех возможных значениях параметров, удовлетворяющих условиям проверяемой гипотезы. Однако чаще всего максимум этой вероятности достигается при значениях параметров, лежащих на границе между гипотезой **H** и альтернативой **K** (именно поэтому утверждения о распределении статистик даются в предположениях этой граничной точки).

В дальнейшем, наряду с видом критической области, мы будем приводить выражение для критического уровня значимости.

Рассмотренный критерий проверки гипотезы **H** : $\rho \leq \rho_0$ основан на статистике $T = \sqrt{n}(r_n - \rho_0)/\hat{\tau}_\rho$, имеющей при $\rho = \rho_0$ (приблизительно) нормальное $(0, 1)$ распределение. Поэтому для значения $T = t_{\text{эксп}}$

$$\alpha_{\text{кр}} = \mathbf{P} \{ T > t_{\text{эксп}} \mid \rho_0 \} \approx 1 - \Phi(t_{\text{эксп}}).$$

§ 9. В предположении независимости величин X_1, X_2 значение коэффициента корреляции $\rho = 0$, а асимптотическая дисперсия (I.2, стр.10)

$$\begin{aligned} \tau_\rho^2 &= \frac{a_{22}}{\sigma_1^2 \sigma_2^2} = \frac{\mathbf{E}(X_1 - \mu_1)^2 (X_2 - \mu_2)^2}{\sigma_1^2 \sigma_2^2} = \left(\begin{array}{c} \text{в силу} \\ \text{независимости} \end{array} \right) \\ &= \frac{\mathbf{E}(X_1 - \mu_1)^2 \mathbf{E}(X_2 - \mu_2)^2}{\sigma_1^2 \sigma_2^2} = 1. \end{aligned}$$

Следовательно, проверка самой популярной на практике гипотезы независимости может быть основана на асимптотически нормальной $\mathcal{N}(0, 1)$ статистике $\sqrt{n} r_n$. Другими словами, при проверке гипотезы независимости критический уровень значимости равен

$$\alpha_{\text{кр}} = \mathbf{P} \{ |r_n| > |r_{\text{эксп}}| \} \approx 2 \left[1 - \Phi \left(\sqrt{n} |r_{\text{эксп}}| \right) \right],$$

где $r_{\text{эксп}}$ – выборочное значение коэффициента корреляции Пирсона.

§ 10. Построение статистических критериев для других характеристик многомерного распределения можно осуществить аналогичным образом. В частности, в теореме I.10 на стр.17 приведено утверждение относительно асимптотического распределения выборочного коэффициента регрессии.

§ 11. Как показывает практика, использование описанных здесь асимптотических методов допустимо только при весьма больших объемах выборок (> 100). В противном случае надежность выводов резко снижается.

Оказывается, для нормальных экспериментальных данных можно построить асимптотический критерий, имеющий низкую вычислительную погрешность уже при достаточно малых объемах выборок.

2.2 Асимптотические выводы в нормальной модели.

Преобразование Фишера

Предположим, что выборочные данные представляют собой независимые реализации многомерных нормальных случайных векторов с неизвестными параметрами. В нормальной вероятностной модели выражения для теоретических моментов выборочных характеристик имеют более простой вид.

Лемма I.3. Если $(X_1, X_2) \sim \mathcal{N}_2(\mu_1, \mu_2 | \sigma_1^2, \sigma_2^2, \rho)$, то

$$\begin{aligned} a_{40} &= 3\sigma_1^4, \quad a_{04} = 3\sigma_2^4, \quad a_{22} = (2\rho^2 + 1)\sigma_1^2\sigma_2^2, \\ a_{13} &= 3\rho\sigma_1\sigma_2^3, \quad a_{31} = 3\rho\sigma_1^3\sigma_2. \end{aligned}$$

Подставляя эти значения в формулу (I.2, стр.10), получаем следующее утверждение об асимптотическом распределении выборочного коэффициента корреляции.

Теорема I.7. Пусть r_n – выборочный полный коэффициент корреляции, построенный по выборке объема n из двумерной нормальной генеральной совокупности с истинным коэффициентом корреляции ρ . Тогда

$$\sqrt{n}(r_n - \rho) \rightsquigarrow \mathcal{N}_1(0, (1 - \rho^2)^2). \quad (\text{I.3})$$

Преобразование Р.Фишера

В соответствии с дельта-методом, асимптотическая дисперсия любой дифференцируемой функции $h(r_n)$ от выборочного коэффициента корреляции равна $(\dot{h}(\rho))^2(1 - \rho^2)^2$. Если выбрать функцию h так, чтобы

$$\dot{h}(\rho) = \frac{dh(\rho)}{d\rho} = \frac{1}{1 - \rho^2},$$

то асимптотическая дисперсия будет равна единице. Функция

$$h(\rho) = \operatorname{arcth}(\rho) = \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho},$$

удовлетворяющая этому дифференциальному уравнению, называется z-преобразованием Фишера, поскольку введена именно Р.Фишером, и чаще всего обозначается латинской буквой „z“, если это преобразование применяется к выборочному коэффициенту корреляции r_n , или греческой буквой „ ζ “, если оно применяется к истинному значению коэффициента корреляции ρ .

Теорема I.8. Пусть r_n – выборочный полный коэффициент корреляции, построенный по выборке объема n из двумерной нормальной генеральной совокупности с истинным коэффициентом корреляции ρ . Тогда преобразование Фишера

$$Z_n = \operatorname{arcth}(r_n) = \frac{1}{2} \ln \frac{1 + r_n}{1 - r_n} \rightsquigarrow \mathcal{N}_1 \left(\zeta, \frac{1}{n} \right) -$$

асимптотически нормально $\mathcal{N}_1(\zeta, 1/n)$ с $\zeta = \operatorname{arcth}(\rho)$.

§ 12. Специально проведенные сравнительные исследования показали, что точность асимптотического приближения распределения статистики Z_n увеличивается, если заменить здесь дисперсию $1/n$ на $1/(n - 3)$. В этом случае уже при объеме выборки $n = 20$ ошибка составляет величину порядка 10^{-3} , что более чем достаточно для практических нужд.

Более подробными вычислениями можно доказать справедливость следующей теоремы.

Теорема I.9. Пусть Z_n – преобразование Фишера для коэффициента корреляции r_n , построенного по нормальной выборке объема n , ζ – преобразование Фишера для истинного коэффициента корреляции ρ ,

$$\mu_z = \zeta + \frac{\rho}{2(n-3)} \left[1 - \frac{3-\rho^2}{4(n-3)} \right], \quad \tau_z^2 = \left[1 - \frac{\rho^2}{2(n-3)} \right].$$

Тогда при $n \rightarrow \infty$

$$\sqrt{n-3} \frac{Z_n - \mu_z}{\tau_z} \rightsquigarrow \mathcal{N}_1(0, 1).$$

Проверка гипотез и доверительные интервалы для ρ

Применим утверждения последних двух теорем для решения задач проверки гипотез о коэффициенте корреляции ρ и построении доверительных интервалов для ρ .

Пусть

- 1) $z = \frac{1}{2} \ln \frac{1+r}{1-r}$ – преобразование Фишера для выборочного полного коэффициента корреляции r , построенного по выборке объема n из нормальной совокупности;
- 2) $\zeta_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0}$ – преобразование Фишера для граничной точки ρ_0 ;
- 3) $\mu_{z0} = \zeta_0 + \frac{\rho_0}{2(n-3)} \left[1 - \frac{3-\rho_0^2}{4(n-3)} \right], \quad \tau_{z0}^2 = 1 - \frac{\rho_0^2}{2(n-3)}$
– асимптотические среднее и дисперсия статистики Z ;
- 4) $\mathcal{N}^{(\frac{\alpha}{2})}$ – верхняя $\frac{\alpha}{2}$ -квантиль нормального распределения.

Тогда

а) при проверке гипотезы $\mathbf{H} : \rho \geq \rho_0$ о коэффициенте корреляции нормальной совокупности критический уровень значимости равен

$$\alpha_{\text{кр}} = \mathbf{P}\{Z_n < z\} \approx \Phi \left(\frac{z - \mu_{z0}}{\tau_{z0}} \sqrt{n-3} \right);$$

б) интервал

$$\text{th} \left(z - \frac{\mathcal{N}(\frac{\alpha}{2})}{\sqrt{n-3}} \right) \leq \rho \leq \text{th} \left(z + \frac{\mathcal{N}(\frac{\alpha}{2})}{\sqrt{n-3}} \right) \quad (\text{I.4})$$

задает асимптотический $(1 - \alpha)$ -доверительный интервал для истинного коэффициента корреляции ρ .

З 13. Обычно при вычислении среднего и дисперсии ограничиваются только первыми слагаемыми, как это и сделано при построении доверительного интервала. Применение уточненных формул с подставленной оценкой $\hat{\rho} = r$ может привести, как ни странно, к ухудшению качества доверительных границ.

2.3 Статистические выводы о среднеквадратической регрессии

– Как жись?

– Нормально $(\mu|\Lambda)!$

Из разговора

Рассмотрим теперь задачу наилучшего прогноза одной случайной величины по наблюдаемым значениям других случайных величин. Если вероятностная модель полностью известна, то регрессия X_1 на X_2 (то есть прогноз значения X_1 по значению $X_2 = x_2$) равна (см.[10, стр.12, (I.2)])

$$x_{1*} = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2) = \mu_1 + \frac{\lambda_{12}}{\lambda_{22}} (x_2 - \mu_2). \quad (\text{I.5})$$

Найдем сначала распределение оценки коэффициента регрессии $\beta = \rho\sigma_1/\sigma_2 = \lambda_{12}/\lambda_{22}$. В качестве **самостоятельной работы** (аналогично доказательствам теорем I.5, I.7) предлагается доказать следующее утверждение.

Теорема I.10. I. Пусть $b_n = l_{12}/s_2^2$ – выборочный коэффициент регрессии, построенный по двумерной выборке объема n с истинным коэффициентом регрессии β и смешанными центральными моментами a_{ml} , $m, l = \overline{0, 4}$. Тогда

$$\sqrt{n}(b_n - \beta) \rightsquigarrow \mathcal{N}_1(0, \tau_\beta^2), \quad n \rightarrow \infty,$$

с асимптотической дисперсией

$$\tau_\beta^2 = \frac{1}{\sigma_2^4}(a_{22} - 2a_{13}\beta + a_{04}\beta^2). \quad (\text{I.6})$$

II. Если величины X_1, X_2 независимы, то

$$\tau_\beta^2 = \frac{\sigma_1^2}{\sigma_2^2}.$$

III. Если выборка получена из двумерного нормального распределения с истинным коэффициентом корреляции ρ , то

$$\tau_\beta^2 = \frac{\sigma_1^2}{\sigma_2^2}(1 - \rho^2). \quad (\text{I.7})$$

§ 14. Формула (I.7) с подставленными в неё оценками неизвестных параметров используется обычно в статистических пакетах для вычисления дисперсии выборочного коэффициента регрессии. В общем случае, конечно, надо бы использовать формулу (I.6), которая, однако, более громоздка и требует вычисления дополнительных моментов, а посему (с учетом общей лени и тайной надежды на то, что всё в этой жизни нормально) и не применима.

При построении регрессии не столь интересен коэффициент регрессии, сколь значение самой функции регрессии. Как видно из формулы (I.5), эта регрессия зависит от четырехмерного вектора неизвестных параметров $\vec{\vartheta} = (\mu_1, \mu_2, \lambda_{22}, \lambda_{12})'$. Вектор оценок этих параметров $\vec{\theta}_n = (X_{1\cdot}, X_{2\cdot}, l_{22}, l_{12})'$ также асимптотически нормален. Его матрица ковариаций может быть найдена также легко, как ковариация вектора вторых моментов в теореме I.4. Для нормальной модели она равна

$$D_{\vartheta} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 2\sigma_2^4 & 2\rho\sigma_1\sigma_2^3 \\ 0 & 0 & 2\rho\sigma_1\sigma_2^3 & (\rho^2 + 1)\sigma_1^2\sigma_2^2 \end{pmatrix}.$$

Теорема I.11. I. Если выборка получена из двумерного нормального распределения $\mathcal{N}_2(\mu_1, \mu_2 | \sigma_1^2, \sigma_2^2, \rho)$, то

$$\sqrt{n}(\vec{\theta}_n - \vec{\vartheta}) \rightsquigarrow \mathcal{N}_4(\vec{0} | D_{\vartheta}), \quad n \rightarrow \infty.$$

II. Оценка

$$\hat{X}_1 = X_{1\bullet} + r_n \frac{s_1}{s_2} (x_2 - X_{2\bullet})$$

наилучшего прогноза сл.в. X_1 по значению сл.в. $X_2 = x_2$ асимптотически нормальна

$$\sqrt{n}(\hat{X}_1 - x_{1*})/\hat{\tau}_1 \rightsquigarrow \mathcal{N}_1(0, 1)$$

со средним, равным истинному наилучшему прогнозу x_{1*} , и дисперсией

$$\frac{1}{n}\hat{\tau}_1^2 = \frac{1}{n}s_1^2(1 - r_n^2) \left(1 + \left(\frac{x_2 - X_{2\bullet}}{s_2} \right)^2 \right). \quad (\text{I.8})$$

В общем случае матрица D_{ϑ} будет зависеть от моментов четвертого порядка (см. доказательство теоремы), и поэтому дисперсия $\hat{\tau}_1^2$ не будет иметь такой законченный вид. Другая причина, по которой на практике применяют исключительно формулу (I.8), состоит в том, что аналогичная, но уже точная, формула дисперсии оценки регрессии получается в ситуации, когда значения переменной X_2 не случайны, а задаются исследователем (см. ниже раздел, посвященный функциональной регрессии).

Доверительный интервал для регрессии

Предыдущую теорему можно использовать для построения доверительного интервала для наилучшего прогноза.

Теорема I.12. Пусть

$$\hat{x}_1 = \hat{x}_1(x) = x_{1\bullet} + r \frac{s_1}{s_2} (x - x_{2\bullet}) \quad -$$

выборочное значение функции регрессии X_1 на X_2 ,

$$\frac{1}{n}\hat{\tau}_1^2 = \frac{1}{n}\hat{\tau}_1^2(x) = \frac{1}{n}s_1^2(1-r^2) \left(1 + \left(\frac{x-x_{2\bullet}}{s_2}\right)^2\right) -$$

дисперсия $\hat{x}_1$, вычисленные при значении $X_2 = x$ по выборке из нормального распределения. Тогда интервал

$$\left(\hat{x}_1 - \frac{\hat{\tau}_1}{\sqrt{n}} \mathcal{N}^{(\frac{\alpha}{2})}; \hat{x}_1 + \frac{\hat{\tau}_1}{\sqrt{n}} \mathcal{N}^{(\frac{\alpha}{2})}\right)$$

задает асимптотический $(1 - \alpha)$ -доверительный интервал для истинного значения регрессии X_1 на $X_2 = x$.

З 15. Имеются некоторые ограничения на применение интервалов подобного рода. Поскольку этот интервал построен для одного значения $X_2 = x$, то его нельзя применять многократно для различных значений x . Проанализируем причину такого запрета с точки зрения общей теории статистического вывода.

Любое $(1 - \alpha)$ -доверительное утверждение (в данном случае доверительный интервал) представляет собой событие A в выборочном пространстве, вероятность которого должна быть больше заданной надежности $(1 - \alpha)$: $\mathbf{P}\{A\} \geq 1 - \alpha$. Например, при $\alpha = 0.1$ это будет означать, что примерно один раз из каждых десяти случаев доверительное утверждение дает неверную рекомендацию. Если мы строим несколько доверительных утверждений $A_1, \dots, A_k$, то должны потребовать, чтобы не просто каждое в отдельности из них имело надежность $(1 - \alpha)$, а общее утверждение, представляющее собой пересечение событий $A_1 \cap \dots \cap A_k$, имело надежность $(1 - \alpha)$. Если этого не потребовать, то очень много шансов (в худшем случае до 65%), что из десяти доверительных утверждений по крайней мере одно обязательно будет ложным.

Если количество ожидаемых доверительных утверждений k фиксированно и мало, причем каждое из них имеет надежность $1 - q$, то, воспользовавшись известным правилом де Моргана (дополнение пересечений равно объединению дополнений), можно заметить, что надежность совместного доверительного утверждения

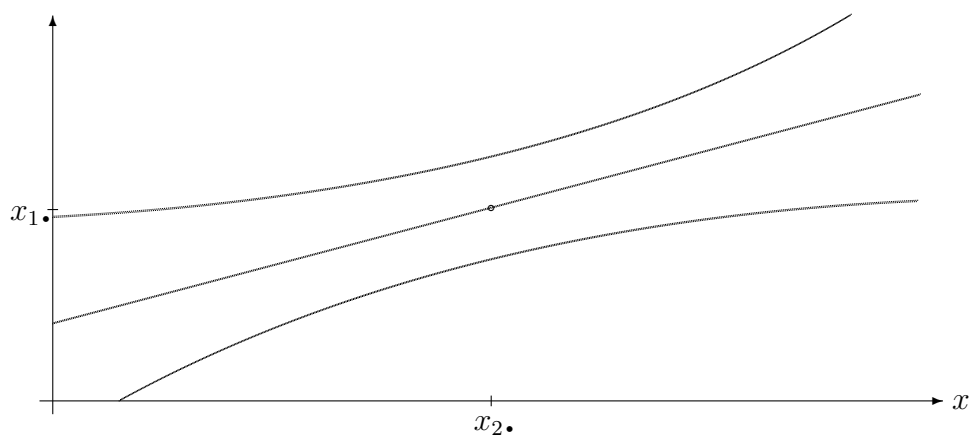
$$\mathbf{P}\left\{\bigcap_{i=1}^k A_i\right\} = 1 - \mathbf{P}\left\{\bigcup_{i=1}^k \bar{A}_i\right\} \geq 1 - \sum_{i=1}^k \mathbf{P}\{\bar{A}_i\} = 1 - \sum_{i=1}^k q = 1 - kq.$$

Таким образом, для того чтобы совместное утверждение имело надежность $(1 - \alpha)$, можно потребовать, чтобы надежность каждого из доверительных утверждений A_i была не меньше $(1 - \frac{\alpha}{k})$. Совместное доверительное утверждение такого типа называется доверительным утверждением Бонферрони. При больших, а тем более бесконечных k (например, когда требуется построить доверительное утверждение для всей линии регрессии), это может оказаться слишком жёстким ограничением.

Замечательно, но здесь можно построить доверительный интервал сразу для всей линии регрессии. Для этого достаточно $\frac{\alpha}{2}$ -квантиль нормального закона $\mathcal{N}(\frac{\alpha}{2})$ заменить на константу $C_n^{(\alpha)} = \sqrt{2 \mathcal{F}_{2,n-2}^{(\alpha)}}$, где $\mathcal{F}_{2,n-2}^{(\alpha)}$ равно верхней α -квантили распределения Фишера с $(2, n - 2)$ степенями свободы. Доверительное утверждение такого сорта называется доверительной полосой Уоркинга-Хотеллинга. Для сравнения приведем несколько значений константы $C_n^{(\alpha)}$ при $\alpha = 0.05$:

$\mathcal{N}(\frac{\alpha}{2})$	$C_{10}^{(\alpha)}$	$C_{20}^{(\alpha)}$	$C_{50}^{(\alpha)}$	$C_{100}^{(\alpha)}$	$C_{500}^{(\alpha)}$
1.96	2.99	2.67	2.53	2.49	2.46

Ниже на рисунке представлен примерный вид графика трех линий (линии регрессии, левой и правой доверительных границ) в зависимости от значения признака $X_2 = x$.



З 16. Часто этот график дополняют ещё парой кривых, „параллельных“ линиям доверительных границ. Предлагается воспринимать этот расширенный интервал как доверительный интервал для возможных значений

сл.в. X_1 (при значении $X_2 = x$). Обосновать такое построение можно простым представлением X_1 в виде суммы её наилучшего прогноза и ошибки – $X_1 = X_{1*} + \varepsilon$. Дисперсия ошибки равна остаточной дисперсии $\sigma_{1*}^2 = \sigma_1^2(1 - \rho^2)$. Таким образом, если доверительный интервал для X_{1*} расширить пропорционально оценке остаточной дисперсии $s_{1*}^2 = s_1^2(1 - r^2)$ (прием, сравнимый с хорошо известным правилом „двух сигм“), то можно получить так называемый толерантный интервал ожидаемых значений сл.в. X_1 . Достигается такое расширение добавлением к асимптотической дисперсии $\frac{\hat{\tau}_1^2}{n}$ величины остаточной дисперсии: $\frac{\hat{\tau}_1^2}{n} \mapsto \frac{\hat{\tau}_1^2}{n} + s_{1*}^2$.

2.4 Функциональная регрессия

Построение доверительного утверждения для значения одного из наблюдаемых признаков – вполне естественная задача в ситуации так называемой функциональной регрессии, когда только прогнозируемый признак случаен, а все остальные признаки задаются исследователем. Например, при сельскохозяйственном эксперименте количество вносимых удобрений, объем полива и другие характеристики задаются самим экспериментатором, а вот результат уже не предсказуем. Вероятностная модель такого эксперимента описывается следующим образом. Результат эксперимента Y связан с p контролируемыми признаками $x_1, \dots, x_p$ соотношением

$$Y = R(x_1, \dots, x_p) + \varepsilon,$$

где R – функция регрессии, ε – случайная ошибка. Предполагается, что ошибка имеет среднее значение, равное нулю, и дисперсию σ^2 , одинаковую для всех входных значений $x_1, \dots, x_p$. Признаки $x_1, \dots, x_p$ называются регрессорами или предикторами, а признак Y откликом. Большинство типов функций регрессии после некоторых преобразований может быть приведено к линейной:

$$R(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

Здесь важна линейность именно по параметрам, а не по предикторам. Например, при изучении влияния расстояния r между двумя предметами на силу их притяжения F исходят из соотношения $F = \beta/r^2$. Полагая

$x = 1/r^2$, получаем стандартную задачу линейного регрессионного анализа. Если ещё учитываются массы обоих предметов, то уравнение связи (регрессии) будет иметь вид $F = \beta_0 m_1^{\beta_1} m_2^{\beta_2} / r^2$. Чтобы свести эту регрессию к линейной, можно обе части уравнения просто прологарифмировать.

§ 17. Подчеркнем существенное различие между двумя типами регрессий. Для задач „структурной“ регрессии со случайными предикторами функция регрессии представляет собой наилучший прогноз значения отклика при наблюдаемых значениях предикторов. При построении „функциональной“ регрессии априори предполагается, что между вектором предикторов и откликом существует функциональная связь выбранного типа, наблюдаемые отклонения от которой обусловлены исключительно ошибками измерения отклика.

В нашем годовом курсе многомерного статистического анализа задачам, связанным с функциональной регрессией, посвящен целый семестр (см. [9]). Правда, там мы изучаем только вопросы дисперсионного анализа, то есть вопросы, относящиеся к проверке гипотез о параметрах линейной регрессии.

Замечательным фактом является то, что оценки коэффициентов регрессии и связанных с ними характеристик для функциональной регрессии почти полностью совпадают с приведенными нами выше, а их распределения находятся гораздо проще. Если предположить, что ошибка ε имеет нормальное распределение, то большинство оценок будет иметь стандартные распределения (хи-квадрат, Стьюдента, Фишера) при конечном объеме выборки, а не только при $n \rightarrow \infty$. Для примера рассмотрим задачу построения доверительного интервала для коэффициента регрессии β_1 при одном регрессоре.

Пусть в эксперименте наблюдается n значений отклика $y_1, \dots, y_n$ при фиксированных значениях $x_{11}, \dots, x_{1n}$ предиктора X_1 . Предполагается, что отклик связан с предикторной переменной равенством

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i,$$

где случайные ошибки ε_i суть независимые реализации нормальной слу-

чайной величины со средним 0 и дисперсией σ^2 . Таким образом, случайность присуща только отклику y , но не предиктору X_1 .

Оценки параметров β_0^* и β_1^* находятся из условия минимизации расстояния между вектором откликов и вектором значений функции регрессии

$$(\beta_0^*, \beta_1^*) = \arg \min_{b_0, b_1} \sum_{i=1}^n (y_i - (b_0 + b_1 x_{1i}))^2.$$

Эти оценки будут также оценками максимального правдоподобия и совпадают с построенными нами ранее оценками параметров линейной регрессии

$$\beta_1^* = l_{xy}/s_x^2, \quad \beta_0^* = y_{\cdot} - \beta_1^* x_{1\cdot},$$

где s_x^2 – дисперсия x -ов, l_{xy} – ковариация (x, y) .

По сложившейся традиции, для функциональной регрессии принято записывать формулы для параметров регрессии в несколько иной форме (см., например, [9]). Для этого введем матрицу

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{1n} & \dots & x_{1n} \end{pmatrix},$$

называемую матрицей плана. В матричной форме вероятностная модель может быть записана в виде

$$\vec{y} = \mathbb{X}' \vec{\beta} + \vec{\varepsilon}.$$

Вектор оценок равен

$$\vec{\beta}^* = \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} = \mathbb{S}^{-1} \mathbb{X} \vec{y},$$

где $\mathbb{S} = \mathbb{X} \mathbb{X}'$ – информационная матрица плана.

Отсюда уже легко получается, что оценки (β_0^*, β_1^*)

а) несмещены: $\mathbf{E} \vec{\beta}^* = \vec{\beta}$;

б) имеют двумерное нормальное распределение

с матрицей ковариаций $\sigma^2 \mathbb{S}^{-1}$:

$$\vec{\beta}^* \sim \mathcal{N}_2(\vec{\beta} | \sigma^2 \mathbb{S}^{-1}).$$

В привычных для модели со случайными предикторами обозначениях

$$\sigma^2 \mathbb{S}^{-1} = \frac{\sigma^2}{n s_x^2} \begin{pmatrix} s_x^2 + x_{1\cdot}^2 & -x_{1\cdot} \\ -x_{1\cdot} & 1 \end{pmatrix}, \quad (\text{I.9})$$

где x_1 — среднее арифметическое значений предикторов, s_x^2 — их выборочная дисперсия. Таким образом, дисперсия

$$\mathbf{D} \beta_1^* = \frac{\sigma^2}{n s_x^2}.$$

Чтобы построить оценку этой дисперсии, нам необходимо найти оценку для σ^2 . В качестве таковой может быть выбрана соответствующим образом нормированная остаточная дисперсия, которая называется суммой квадратов ошибок и обозначается SS_e :

$$SS_e = \sum_{i=1}^n (y_i - (\beta_0^* + \beta_1^* x_{1i}))^2 = n s_y^2 (1 - r^2),$$

где r — коэффициент корреляции Пирсона для пар данных $(y_1, x_{11}), \dots, (y_n, x_{1n})$. Последняя формула приведена здесь снова, чтобы подчеркнуть связь между двумя принятыми способами записи. Можно показать, что SS_e / σ^2 стохастически не зависит от вектора оценок и имеет хи-квадрат распределение с $(n - 2)$ -мя степенями свободы (см. [9, стр.15]). Отсюда можно сделать два далеко идущих вывода.

1. Оценка дисперсии $\hat{\sigma}^2 = \frac{SS_e}{n - 2}$ несмещена.
2. Отношение

$$\frac{\sqrt{n} s_x^2 (\beta_1^* - \beta_1)}{\sqrt{SS_e}} \sqrt{n - 2} \sim \mathcal{S}_{n-2},$$

то есть имеет распределение Стьюдента с $(n - 2)$ -мя степенями свободы. Таким образом, границы $(1 - \alpha)$ -доверительного интервала для β_1 (с учетом второго способа записи SS_e) равны

$$\beta_1^* \pm \frac{s_y}{s_x} \sqrt{1 - r^2} \frac{t_{n-2}^{(\frac{\alpha}{2})}}{\sqrt{n - 2}},$$

где $t_{n-2}^{(\frac{\alpha}{2})}$ — верхняя $\frac{\alpha}{2}$ -квантиль распределения Стьюдента с $(n - 2)$ -мя степенями свободы.

Заметим, что асимптотический доверительный интервал, построенный на основе Теоремы I.6 и формулы (I.7, стр.18), почти полностью совпадает с полученным здесь интервалом. Отличия заключаются лишь в том, что делитель $\sqrt{n - 2}$ заменяется на $\sqrt{n}$, а константа $t_{n-2}^{(\frac{\alpha}{2})}$ выбирается как квантиль стандартного нормального закона. Однако оба эти отличия не

играют существенной роли при больших значениях n , каковые и предполагаются в любых асимптотических утверждениях. Этот факт объясняет причину, по которой дисперсию оценки параметра регрессии β_1 , а вместе с ней и доверительные границы для β_1 , вычисляют всегда по формулам, приведенным в этом разделе.

2.5 Регрессия в многомерном случае

Связь между характеристиками оценок для функциональной регрессии и соответствующими характеристиками оценок для модели со случайными предикторами может быть перенесена и на многомерный случай. Например, если изучается влияние на отклик $Y = X_1$ вектора предикторов $\vec{X} = (X_2, \dots, X_p)'$, то в схеме с n фиксированными наборами значений этих предикторов (схема функциональной регрессии) оценка функции регрессии во входной точке $\vec{X} = \vec{x}_0 = (x_{20}, \dots, x_{p0})'$ равна

$$\hat{\eta}(\vec{x}_0) = (1, \vec{x}_0') \mathbb{S}^{-1} \mathbb{X} \vec{y}, \quad (\text{I.10})$$

где $(1, \vec{x}_0') = (1, x_{20}, \dots, x_{p0})$ – строка входных точек с добавленной первой компонентой, равной единице,

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix} -$$

матрица плана статистического эксперимента,

$\vec{y}$ – вектор n откликов, полученных при входных значениях предикторов, определяемых матрицей плана $\mathbb{X}$,

$\mathbb{S} = \mathbb{X}\mathbb{X}'$ – информационная матрица.

Дисперсия этой оценки равна

$$\begin{aligned} \mathbf{D} \hat{\eta}(\vec{x}_0) &= (1, \vec{x}_0') \mathbb{S}^{-1} \mathbb{X} \mathbf{D} \vec{y} \mathbb{X}' \mathbb{S}^{-1} (1, \vec{x}_0')' = \\ &= \sigma^2 (1, \vec{x}_0') \mathbb{S}^{-1} \mathbb{X} \mathbb{X}' \mathbb{S}^{-1} (1, \vec{x}_0')' = \\ &= \sigma^2 (1, \vec{x}_0') \mathbb{S}^{-1} (1, \vec{x}_0')', \end{aligned}$$

а её несмещенная оценка

$$\frac{1}{n-p} \tau_{\eta}^2(\vec{x}_0) = \frac{1}{n-p} S S_e (1, \vec{x}_0') \mathbb{S}^{-1} (1, \vec{x}_0')'. \quad (\text{I.11})$$

З 18. Положив в этой формуле строку $(1, \vec{x}_0') = (1, 0, \dots, 0)$, получим дисперсию оценки свободного члена регрессии

$$\frac{1}{n-p} \tau_C^2 = \frac{S S_e}{n-p} \tilde{s}_{11}, \quad (\text{I.12})$$

где $\tilde{s}_{11}$ – диагональный элемент обратной матрицы $\mathbb{S}^{-1}$.

Если теперь обратиться к задаче построения наилучшего прогноза значения случайной величины X_1 по значениям случайных же величин $X_2, \dots, X_p$, то наилучшая оценка этого прогноза вычисляется по той же формуле (I.10) с вектором откликов $\vec{y} = (x_{11}, \dots, x_{1n})'$.

Поскольку матрица плана теперь случайна, то можно утверждать только, что асимптотическая дисперсия оценки регрессии будет равна

$$\frac{1}{n-p} \hat{\tau}_{\eta}^2(\vec{x}_0) = \frac{s_1^2(1 - r_{1*}^2)}{n-p} (1, \vec{x}_0') \left(\frac{1}{n} \mathbb{S} \right)^{-1} (1, \vec{x}_0')', \quad (\text{I.13})$$

где s_1^2 – выборочная дисперсия компоненты X_1 ,

r_{1*} – выборочный множественный коэффициент корреляции X_1 с $(X_2, \dots, X_p)$,

$\vec{x}_0 = (x_{20}, \dots, x_{p0})'$ – входные значения $X_2, \dots, X_p$.

Информационную матрицу плана $\mathbb{S}$ легче всего вычислять, исходя из её определения $\mathbb{S} = \mathbb{X} \mathbb{X}'$. Если же это невозможно (программа Excel, например, оперирует только с матрицами, имеющими не более 256 столбцов), то можно воспользоваться следующей формулой, связывающей её с матрицей ковариаций L вектора всех наблюдаемых признаков $(X_1, \dots, X_p)'$:

$$\mathbb{S} = n \left[L^{(0)} + (1, \vec{x}.)' (1, \vec{x}.) \right], \quad (\text{I.14})$$

где $(1, \vec{x}.) = (1, x_{2.}, \dots, x_{p.})$ – строка выборочных средних признаков $X_2, \dots, X_p$ с добавленной единицей на первом месте, а матрица $L^{(0)}$ отличается от матрицы L только тем, что её первые строка и столбец равны нулю.

Асимптотический $(1 - \alpha)$ -доверительный интервал для наилучшего прогноза значения случайного признака X_1 при фиксированных значениях признаков $X_2 = x_{20}, \dots, X_p = x_{p0}$ задается границами

$$\hat{\eta}(\vec{x}_0) \pm \frac{\tau_{\eta}(\vec{x}_0)}{\sqrt{n-p}} \mathcal{N}^{(\frac{\alpha}{2})}, \quad (\text{I.15})$$

с константой $\mathcal{N}^{(\frac{\alpha}{2})}$, равной, в соответствии с утверждением об асимптотической нормальности оценки $\hat{\eta}$, верхней $\frac{\alpha}{2}$ -квантили нормального закона.

Если требуется построить доверительное множество сразу для всей линии регрессии, то, заменив константу $\mathcal{N}^{(\frac{\alpha}{2})}$ на константу

$$C_n^{(\alpha)} = \sqrt{p \mathcal{F}_{p, n-p}^{(\alpha)}},$$

получим p -мерный аналог доверительной полосы Уоркинга-Хотеллинга (см. стр.21). Теоретическое обоснование надежности этой „полосы“ можно найти в монографии [4, стр. 487-493].

§ 3. Точные распределения выборочных моментов в нормальной модели. Теорема Уишарта

Многократное прочтение инструкции
если и не поможет собрать сложный
прибор, зато точно научит витиевато
выражаться.

Для нормальной вероятностной модели можно найти точное распределение оценок некоторых из рассматриваемых нами параметров. В одномерном случае распределение выборочной ковариации $n \vec{\mathcal{L}} (= n s_1^2)$ задается хи-квадрат распределением с $\nu = (n-1)$ степенями свободы. На многомерный случай это распределение обобщает распределение Уишарта.

О п р е д е л е н и е 1. Пусть $\mathbf{U}^{(p,p)}$ – симметричная случайная матрица, $\vec{\mathfrak{U}} = \frac{p(p+1)}{2}$ -мерный вектор определяющих её случайных величин. Распределением Уишарта $\mathbb{W}_p(\Lambda, \nu)$ вектора $\vec{\mathfrak{U}}$ (матрицы $\mathbf{U}$) называется

распределение с плотностью

$$f(\vec{u} | \Lambda, \nu) = \begin{cases} C_{p\nu} |U|^{\frac{\nu-p-1}{2}} \exp \left\{ -\frac{1}{2} \operatorname{tr}(U\Lambda^{-1}) \right\} & , \text{ если } U > 0, \\ 0 & , \text{ если } U \not> 0, \end{cases}$$

где

$$C_{p\nu} = \left(\frac{1}{\pi} \right)^{\frac{p(p-1)}{4}} \left(\frac{2^{-p}}{|\Lambda|} \right)^{\frac{\nu}{2}} \prod_{j=1}^p \frac{1}{\Gamma(\frac{\nu-j+1}{2})}.$$

Рассмотрим теперь вектор выборочных средних $\vec{X}_\bullet = (X_{1\bullet}, \dots, X_{p\bullet})'$ и вектор всех вторых моментов $\vec{\mathcal{L}} = (l_{11}, l_{12}, \dots, l_{pp})$ размерности $p(p+1)/2$, элементы которого образуют матрицу вторых моментов L .

Теорема I.13. Если выборка объема n получена из p -мерного нормального распределения $\mathcal{N}_p(\vec{\mu}, \Lambda)$, то

- (i) совокупности переменных $\vec{X}_\bullet$ и $\vec{\mathcal{L}}$ независимы;
- (ii) вектор $\vec{X}_\bullet \sim \mathcal{N}_p(\vec{\mu}, \frac{1}{n}\Lambda)$;
- (iii) вектор $n\vec{\mathcal{L}} \sim \mathbb{W}_p(\Lambda, \nu)$ с параметром $\nu = n - 1$.

3.1 Моменты обобщенной дисперсии

Воспользуемся выражением плотности распределения Уишарта для отыскания величины первого момента и теоретической дисперсии оценки обобщенной дисперсии $\hat{\mathfrak{S}}^2 = |L|$.

Теорема I.14. Истинный момент $\hat{\mathfrak{S}}^2$ порядка m равен

$$\mathbf{E}|L|^m = \left(\frac{2^p \mathfrak{S}^2}{n^p} \right)^m \prod_{j=1}^p \frac{\Gamma(\frac{n-j}{2} + m)}{\Gamma(\frac{n-j}{2})}.$$

В частности, среднее значение

$$\mathbf{E}|L| = \mathfrak{S}^2 \left(1 - \frac{1}{n} \right) \left(1 - \frac{2}{n} \right) \cdots \left(1 - \frac{p}{n} \right),$$

дисперсия

$$\mathbf{D}|L| = \frac{\mathfrak{S}^4}{n} \left(2 - \frac{p-1}{n}\right) \prod_{j=1}^p \left(1 - \frac{j}{n}\right) \prod_{l=1}^{p-2} \left(1 - \frac{l}{n}\right).$$

Как и ожидалось, оценка $|L|$ смещена. Её можно исправить, поделив на мешающую константу.

В частном случае $p = 1$ обобщенная дисперсия равна дисперсии X_1 , а полученное здесь выражение для $\mathbf{D}|L|$ совпадает с формулой для дисперсии s^2 , приведенной в замечании $\mathfrak{Z}4$, стр.7.

Следствие 1. При объеме выборки $n \rightarrow \infty$ оценка обобщенной дисперсии $\hat{\mathfrak{S}}^2 = |L|$

а) асимптотически несмещена: $\mathbf{E}|L| \rightarrow \mathfrak{S}^2$;

б) состоятельна в средне-квадратическом: $\mathbf{E}(|L| - \mathfrak{S}^2)^2 \rightarrow 0$.

$\mathfrak{Z}19$. Свойство б) влечет слабую состоятельность:

$$\mathbf{P}\text{-}\lim_n |L| = \mathfrak{S}^2.$$

3.2 Распределение выборочных коэффициентов корреляции

Найдем теперь распределение выборочного полного коэффициента корреляции между двумя случайными величинами (например, между X_1 и X_2 .) Обозначим истинный коэффициент корреляции ρ_{12} через ρ , а соответствующий ему выборочный аналог r_{12} через r .

Теорема I.15. Распределение выборочного коэффициента корреляции r зависит только от истинного коэффициента корреляции ρ и от объема выборки n . Функция плотности r равна (при $r \in [-1; 1]$)

$$f_n(r|\rho) = \frac{2^{n-3} (1 - \rho^2)^{\frac{n-1}{2}}}{\pi(n-3)!} (1 - r^2)^{\frac{n-4}{2}} * \sum_{j=0}^{\infty} \left(\Gamma^2 \left(\frac{n+j-1}{2} \right) \frac{(2\rho r)^j}{j!} \right).$$

Для практических целей сама функция плотности не представляет большого интереса – важно уметь вычислять функцию распределения $F_n(x | \rho) = \int_{-1}^x f_n(r | \rho) dr$, которую в дальнейшем использовать для проверки гипотез об истинном коэффициенте корреляции, а также для построения доверительных интервалов.

Почти очевидно, что функция распределения F_n монотонно убывает по параметру ρ , поскольку с увеличением истинного значения коэффициента корреляции ρ увеличивается ожидаемое значение выборочного коэффициента корреляции r , а значит, уменьшается вероятность малых значений, то есть уменьшается вероятность события $\{r < x\}$. Для таких функций распределения справедлива

Теорема I.16. При значении выборочного коэффициента корреляции $r = r_{\text{эксп}}$

(i) нижняя $(1 - \alpha)$ -доверительная граница $\underline{\rho}$ для неизвестного параметра ρ находится как решение уравнения

$$F_n(r_{\text{эксп}} | \underline{\rho}) = 1 - \alpha;$$

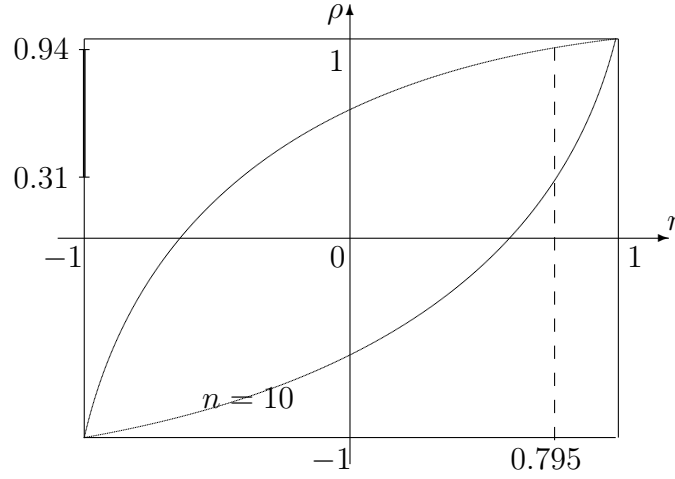
(ii) верхняя граница $\bar{\rho}$ находится как решение уравнения

$$1 - F_n(r_{\text{эксп}} | \bar{\rho}) = 1 - \alpha.$$

В эпоху слабого развития компьютерной техники распределение выборочного коэффициента корреляции можно было найти только в табличной форме. Такого рода таблицы дополнялись, кроме всего прочего, графиками доверительных интервалов. На следующем рисунке, в качестве примера, приведен способ построения 95%-доверительного интервала для ρ при выборочном значении $r = 0.795$ и объеме выборки $n = 10$. Таким образом, 95%-ый доверительный интервал для ρ равен

$$(\underline{\rho}; \bar{\rho}) = (0.31; 0.94).$$

Приближенный „фишеровский“ доверительный интервал (I.4, стр.17) равен $(0.332; 0.950)$, то есть почти не отличается от точного.



З 20. Полученные в этом примере значения для границ свидетельствуют об очень плохих качествах коэффициента r как оценки ρ при малом объеме выборки: столь высокое значение выборочного коэффициента привело к слишком широкому доверительному интервалу. Многие специалисты рекомендуют применять методы корреляционного анализа при объеме выборки не менее 50-100. Так, при объеме выборки $n = 100$ точный доверительный интервал для ρ при выборочном значении $r = 0.795$ равен $(0.708; 0.857)$.

К сожалению, ряд, определяющий функцию плотности коэффициента корреляции Пирсона, сходится очень медленно. Г.Хотеллинг предложил другую форму записи функции плотности

$$f_n(r|\rho) = \frac{(n-2) \Gamma(n-1) (1-\rho^2)^{\frac{n-1}{2}}}{\sqrt{2\pi} \Gamma(n-\frac{1}{2})} \frac{(1-r^2)^{\frac{n-4}{2}}}{(1-\rho r)^{n-\frac{3}{2}}} \cdot {}_2F_1\left(\frac{1+\rho r}{2} \middle| \frac{1}{2}; \frac{1}{2}; n-\frac{1}{2}\right), \quad (\text{I.16})$$

где $\mathcal{G}$ – конфлуэнтная гипергеометрическая функция

$$\mathcal{G}(x|a; b; c) = 1 + \sum_{k=1}^{\infty} \frac{\Gamma(a+k)}{\Gamma(a)} \frac{\Gamma(b+k)}{\Gamma(b)} \frac{\Gamma(c)}{\Gamma(c+k)} \frac{x^k}{k!}.$$

Этот ряд сходится намного быстрее, и для вычисления функции распределения выборочной корреляции следует использовать функцию плотности именно в этой форме.

З 21. Очень часто вместо оценки r предлагается её исправленный на несмещенность вариант

$$\hat{\rho} = r + \frac{r(1 - r^2)}{2n - 8}.$$

Действительно, как показывают прямые вычисления, математическое ожидание исправленной оценки очень близко к истинному значению ρ . Кроме того, при $|\rho| > 0,6$ исправленная оценка имеет и более низкую среднеквадратическую ошибку. Однако, при $|\rho| < 0,6$, а в особенности при $\rho \approx 0$, с точки зрения среднеквадратического риска более предпочтительна оценка r . В этом случае выгоды от несмещенности сравнимы с потерями в точности. Следует семь раз подумать, так ли уж нужна эта „пресловутая“ несмещенность.

Рассматриваемая здесь исправленная оценка представляет собой усе-
чение (до двух членов) разложения в ряд несмещенной оценки ρ

$$r^* = r \mathcal{G} \left(1 - r^2 \left| \frac{1}{2}; \frac{1}{2}; \frac{n-2}{2} \right. \right),$$

приведенной в [4, стр. 393]. Здесь снова функция $\mathcal{G}$ есть гипергеометрическая функция (I.16). Было бы интересно исследовать качество этой оценки с точки зрения среднеквадратического риска (тема **курсовой работы**).

На практике чаще всего требуется проверить гипотезу о независимости признаков.

Теорема I.17. Если выборка получена из p -мерного нормального распределения с независимыми компонентами ($\rho_{kl} = 0, \forall k, l = \overline{1, p}, k \neq l$), то функция плотности вектора выборочных коэффициентов корреляции $\vec{\mathcal{R}}$ равна

$$f(\vec{\mathcal{R}}) = \frac{1}{\pi^{\frac{p(p-1)}{4}}} \prod_{j=1}^p \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-j}{2})} |R|^{\frac{n-p-2}{2}}, \quad R > 0,$$

где R – матрица корреляций, соответствующая вектору $\vec{\mathcal{R}}$.

3.3 Преобразование Стьюдента

В двумерном случае Стьюдент предложил преобразование выборочного коэффициента корреляции, распределение которого имеет „привычный“ для прикладной статистики вид.

Теорема I.18. Если r_n – выборочный коэффициент корреляции, вычисляемый по выборке объема n из двумерного нормального распределения с истинным коэффициентом корреляции $\rho = 0$, то преобразование Стьюдента

$$T = \sqrt{n-2} \frac{r_n}{\sqrt{1-r_n^2}} \quad (\text{I.17})$$

имеет распределение Стьюдента $\mathcal{S}_{n-2}$ с $(n-2)$ -мя степенями свободы.

Для проверки гипотезы о равенстве нулю истинного коэффициента корреляции используется именно преобразование Стьюдента.

Следствие 2. Пусть $t_{\text{эксп}}$ – преобразование Стьюдента для выборочного коэффициента корреляции, построенного по выборке из двумерного нормального распределения. Тогда критический уровень значимости для критерия проверки гипотезы $\mathbf{H} : \rho = 0$ при альтернативе $\mathbf{K} : |\rho| > 0$ равен

$$\alpha_{\text{кр}} = \mathbf{P}\{|T| > |t_{\text{эксп}}|\} = 2[1 - \mathcal{S}_{n-2}(|t_{\text{эксп}}|)]. \quad (\text{I.18})$$

Немного о частной корреляции

Не вдаваясь в подробности доказательства, можно сформулировать следующее утверждение (см., например, [1, стр.119]).

Теорема I.19. Пусть $F_n(r \mid \rho_{12})$ – функция распределения полного выборочного коэффициента корреляции r_{12} между с.в. X_1, X_2 , построенного по выборке объема n из нормального распределения с истинным коэффициентом корреляции ρ_{12} . Тогда функция распределения выборочного частного коэффициента корреляции $r_{12*(3...p)}$ между с.в. X_1, X_2 за вычетом влияния с.в. $X_3, \dots, X_p$ с истинным частным коэффициентом корреляции $\rho_{12*(3...p)}$ равна

$$F_{n-(p-2)}(r \mid \rho_{12*(3...p)}),$$

то есть отличается от функции распределения r_{12} заменой объема выборки n на $n - (p - 2)$.

Из этого факта следует, что статистические выводы относительно частного коэффициента корреляции строятся аналогично выводам относительно полного коэффициента корреляции. Например, преобразование Стьюдента для частного коэффициента корреляции равно

$$t = \sqrt{n-p} \frac{r_{12*}}{\sqrt{1-r_{12*}^2}},$$

и в предположении равенства нулю истинного частного коэффициента корреляции оно имеет распределение Стьюдента с $(n - p)$ степенями свободы. Отсюда уже легко построить критерий проверки гипотезы $\mathbf{H} : \rho_{12*} = 0$.

Аналогично (с помощью соответствующего преобразования Фишера) строятся доверительные интервалы для ρ_{12*} .

§ 22. Такое замечательное соотношение между полным и частным коэффициентами корреляции имеет место только для нормальной модели. Выводы о ρ_{12*} в общем случае будут зависеть уже от всей неизвестной матрицы ковариаций. В связи с этим восхищает „осторожность“ разработчиков пакетов статистической обработки, в которых критический уровень значимости вычисляется только для полного (Пирсоновского) коэффициента корреляции, но не для частного коэффициента корреляции.

3.4 Проверка независимости групп случайных векторов

Часто на практике возникает задача проверки гипотезы независимости двух или более групп случайных признаков (например, независимость всех измеряемых показателей здоровья от всех проводимых лечебных мероприятий). В целях сокращения записи рассмотрим только случай с двумя группами. Итак, пусть наблюдаемый вектор $\vec{X} = (\vec{U}', \vec{W}')'$ представляет собой совокупность двух векторов $\vec{U}$ и $\vec{W}$. Требуется проверить гипотезу

$$\mathbf{H} : \vec{U} \text{ не зависит от } \vec{W}.$$

Матрица Λ истинных коэффициентов ковариации вектора $\vec{X}$ может быть разбита на четыре блока

$$\Lambda = \mathbf{Cov}(\vec{X}) = \begin{pmatrix} \Lambda_U & \Lambda_{UW} \\ \Lambda_{WU} & \Lambda_W \end{pmatrix},$$

где $\Lambda_U = \mathbf{Cov}(\vec{U})$ – матрица ковариаций вектора $\vec{U}$,

$\Lambda_W = \mathbf{Cov}(\vec{W})$ – матрица ковариаций вектора $\vec{W}$,

$\Lambda_{UW} = \mathbf{Cov}(\vec{U}, \vec{W}) = \Lambda'_{WU}$ – матрицы ковариаций векторов $\vec{U}, \vec{W}$.

Если гипотеза $\mathbf{H}$ верна, тогда все перекрестные ковариации равны нулю и матрица

$$\Lambda = \begin{pmatrix} \Lambda_U & \mathbf{0} \\ \mathbf{0} & \Lambda_W \end{pmatrix}.$$

Построим критерий отношения правдоподобия для проверки этой гипотезы. Для этого найдем оценки максимального правдоподобия для ковариаций в предположениях гипотезы. Если гипотеза верна, то функция правдоподобия может быть представлена в виде произведения функций правдоподобия наблюдений над вектором $\vec{U}$ и вектором $\vec{W}$ соответственно. Поэтому оценки ковариаций векторов $\vec{U}$ и $\vec{W}$ будут совпадать с оценками соответствующих ковариаций всего вектора $\vec{X}$, а максимум правдоподобия в силу (I.1, стр.6) будет равен

$$\max_{\vec{\mu}, \Lambda \in \mathbf{H}} f(\vec{x}^{(n)} | \vec{\mu}, \Lambda) = (2\pi e)^{-\frac{pn}{2}} |L_U|^{-\frac{n}{2}} \cdot |L_W|^{-\frac{n}{2}}.$$

Таким образом, статистика критерия отношения правдоподобия равна

$$T = \frac{\max_{\vec{\mu}, \Lambda \in \mathbf{H}} f(\vec{x}^{(n)} | \vec{\mu}, \Lambda)}{\max_{\vec{\mu}, \Lambda} f(\vec{x}^{(n)} | \vec{\mu}, \Lambda)} = \left[\frac{|L|}{|L_U| |L_W|} \right]^{\frac{n}{2}}.$$

Точное распределение статистики T известно только для некоторых частных случаев (например, для приведенного ниже критерия проверки гипотезы независимости одной компоненты от всех остальных). В монографии [1, с.320-326] приведено асимптотическое разложение распределения статистики отношения правдоподобия T для проверки гипотезы независимости сразу J групп случайных величин, содержащих по $p_1, p_2, \dots, p_J$ компонент ($p_1 + \dots + p_J = p$). Остановимся вкратце на описании главной части этого разложения, несколько видоизменив тестовую статистику.

Пусть $\mathcal{K}_\nu$ – функция хи-квадрат распределения с ν степенями свободы,

L – выборочная матрица ковариаций вектора $\vec{X} = (\vec{U}'_1, \dots, \vec{U}'_J)'$,

L_j – часть этой матрицы, соответствующая подвектору $\vec{U}_j$.

Теорема I.20. Предположим, что совместное распределение наблюдаемого случайного вектора $\vec{X} = (\vec{U}'_1, \dots, \vec{U}'_J)'$ нормально.

I. Статистика критерия отношения правдоподобия для проверки гипотезы независимости J векторов $\vec{U}_1, \dots, \vec{U}_J$ равна

$$\mathfrak{T} = \sum_j \ln(|L_j|) - \ln(|L|).$$

II. Критический уровень значимости критерия отношения правдоподобия для выборочного значения $\mathfrak{T} = \hat{\mathfrak{T}}$ равен

$$\alpha_{кр} = \mathbf{P}\{\mathfrak{T} > \hat{\mathfrak{T}}\} = 1 - \mathcal{K}_\nu(\eta \hat{\mathfrak{T}}) + O(1/\eta^2), \quad (\text{I.19})$$

где $\nu = (p^2 - \sum p_j^2)/2$, $\eta = n - 1.5 - (p^3 - \sum p_j^3)/(6\nu)$.

З 23. Из общей теории асимптотических распределений известно, что умноженный на (-2) логарифм статистики отношения правдоподобия имеет в пределе хи-квадрат распределение с числом степеней свободы, равным разности количеств неизвестных параметров общей модели и модели, задающей гипотезу. В условиях приведенной здесь теоремы общее число параметров равно $p + \frac{1}{2}p(p+1)$ (p средних и $\frac{1}{2}p(p+1)$ ковариаций), а в

предположениях гипотезы $\mathbf{H} - (p + \sum \frac{1}{2}p_j(p_j + 1))$, что объясняет появление числа ν в теореме.

Вместо параметра η можно использовать просто объем выборки n . Приведенный здесь вариант асимптотического утверждения дает более точную аппроксимацию.

В частном случае, когда требуется проверить гипотезу независимости сразу всех компонент наблюдаемого вектора, все $p_j = 1$, а одномерные матрицы $L_j = s_j^2$ совпадают с соответствующими дисперсиями.

Следствие 3. I. Статистика критерия отношения правдоподобия для проверки гипотезы независимости всех компонент вектора $\vec{X}$ зависит от матрицы полных выборочных корреляций R :

$$\mathfrak{T} = -\ln(|R|).$$

II. Критический уровень значимости критерия отношения правдоподобия равен (I.19) с

$$\nu = \frac{p(p-1)}{2}, \quad \eta = n - \frac{2p+11}{6}.$$

3.5 Проверка гипотезы независимости по множественному коэффициенту корреляции

Если при проверке гипотезы независимости двух векторов первый вектор $\vec{U}$ содержит только одну компоненту, то одномерная матрица

$L_U = s_1^2$ – равна выборочной дисперсии компоненты X_1 , а матрица

$L_W = L_{11}$ – есть матрица ковариаций L без первой строки и первого столбца. Следовательно, статистика T может быть записана через множественный коэффициент корреляции r_{1*} (см.[10, стр.11, Теорема I.3]) как

$$T = (1 - r_{1*}^2)^{\frac{n}{2}}.$$

Таким образом, проверка гипотезы независимости одной переменной X_1 от всех остальных компонент наблюдаемого случайного вектора может быть основана на множественном коэффициенте корреляции r_{1*} .

Теорема I.21. Пусть $\hat{r}_{1*}$ – выборочное значение множественного коэффициента корреляции, построенного по выборке объема n из p -мерного нормального распределения, тогда при проверке гипотезы $\mathbf{H} : \rho_{1*} = 0$ об истинном множественном коэффициенте корреляции критический уровень значимости критерия отношения правдоподобия равен

$$\alpha_{кр} = \mathbf{P}\{r_{1*} > \hat{r}_{1*}\} = 1 - \mathcal{F}_{p-1, n-p} \left(\frac{n-p}{p-1} \cdot \frac{\hat{r}_{1*}^2}{1 - \hat{r}_{1*}^2} \right).$$

Идея доказательства теоремы состоит в том, что разность $1 - r_{1*}^2$ равна отношению остаточной дисперсии s_{1*}^2 к полной дисперсии первой компоненты s_1^2 [10, стр.11, Теорема I.3]. В терминах теории дисперсионного анализа дисперсия s_{1*}^2 равна (с точностью до множителя $1/n$) сумме квадратов ошибок SS_e . Из курса дисперсионного анализа известно (см., например, [9, стр.19, Теорема II.4]), что если предикторные переменные не случайны, то отношение

$$\frac{n-p}{p-1} \frac{n s_1^2 - SS_e}{SS_e} \sim \mathcal{F}_{p-1, n-p} \quad -$$

имеет распределение Фишера с параметрами $(p-1, n-p)$.

З 24. При $p = 2$ множественный коэффициент корреляции равен модулю полного коэффициента корреляции, а распределение Фишера $\mathcal{F}_{1, n-2}$ совпадает с распределением квадрата „студентовской“ с.в. Таким образом, при $p = 2$ предложенный здесь критерий совпадает с двусторонним критерием Стьюдента для полного коэффициента корреляции.

В случае, когда истинный коэффициент $\rho_{1*} \neq 0$, распределение квадрата выборочной множественной корреляции r_{1*}^2 зависит только от ρ_{1*} и его функция плотности имеет вид, схожий с функцией плотности полного коэффициента корреляции (I.16, стр.32). Точное утверждение можно найти в монографии [1, стр.134].

Теорема I.22. Пусть r_{1*} – выборочный множественный коэффициент корреляции, построенный по выборке объема n из p -мерного нормального распределения с истинным множественным коэффициентом корреляции ρ_{1*} . Тогда функция плотности r_{1*}^2 равна (при $0 \leq r \leq 1$)

$$f_n^*(r|\rho) = \frac{\Gamma\left(\frac{n-1}{2}\right) (1 - \rho_{1*}^2)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-p}{2}\right) \Gamma\left(\frac{p-1}{2}\right)} r^{\frac{p-3}{2}} (1 - r)^{\frac{n-p-2}{2}} \cdot \mathcal{G}\left(r\rho_{1*}^2 \left| \frac{n-1}{2}; \frac{n-1}{2}; \frac{p-1}{2} \right.\right),$$

где $\mathcal{G}$ – гипергеометрическая функция (стр.32).

З 25. Поскольку при истинном $\rho_{1*} = 0$ функция $\mathcal{G} = 1$, то при справедливости гипотезы независимости с.в. X_1 от всех остальных компонент вектора $\vec{X}$ распределение квадрата выборочного множественного коэффициента корреляции r_{1*}^2 совпадает с бета-распределением с параметрами $(\frac{p-1}{2}, \frac{n-p}{2})$.

§ 4. Доказательства

Если не уверен, что тебя поймут, то
постарайся хотя бы убедить всех в
своем понимании.

Доказательство Теоремы I.1.

Для функции плотности нормального распределения

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^p |\Lambda|}} \exp \left\{ -\frac{1}{2} (\vec{x} - \vec{\mu})' \Lambda^{-1} (\vec{x} - \vec{\mu}) \right\}$$

логарифм функции правдоподобия выборочных данных

$$\begin{aligned} \mathbb{L}(\vec{\mu}, \Lambda | \vec{x}^{(n)}) &= \ln \prod_{i=1}^n f(\vec{x}_i) = \sum_{i=1}^n \ln f(\vec{x}_i) = \\ &= -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda^{-1}| - \frac{1}{2} \sum_{i=1}^n (\vec{x}_i - \vec{\mu})' \Lambda^{-1} (\vec{x}_i - \vec{\mu}) = \\ &= -\frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda^{-1}| - \frac{1}{2} \sum_{i=1}^n \sum_{j,k=1}^p q_{jk} (x_{ji} - \mu_j)(x_{ki} - \mu_k), \end{aligned} \tag{I.20}$$

где q_{jk} - элементы обратной матрицы Λ^{-1} . Производная этой функции по параметру μ_j равна

$$\frac{\partial \mathbb{L}}{\partial \mu_j} = \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^p q_{jk} (x_{ki} - \mu_k) = \frac{n}{2} \sum_{k=1}^p q_{jk} (x_{k\cdot} - \mu_k).$$

Таким образом, система уравнений максимального правдоподобия $\partial \mathbb{L} / \partial \mu_j = 0, j = \overline{1, p}$, может быть записана в матричной форме

$$\Lambda^{-1} (\vec{x}_{\cdot} - \vec{\mu}) = \vec{0}.$$

Поскольку матрица Λ^{-1} не вырождена, то последнее соотношение имеет место только при $\vec{\mu} = \vec{x}_{\cdot}$, что доказывает утверждение теоремы относительно оценок средних нормального распределения.

Для отыскания максимума по элементам матрицы Λ заметим сначала, что в силу взаимной однозначности связи между матрицей Λ и обратной

к ней матрицей Λ^{-1} этот максимум можно искать по компонентам q_{jk} матрицы Λ^{-1} . Кроме того, будем считать элементы q_{jk} и q_{kj} не обязательно равными.

Найдем производную по параметру q_{jk} . Для этого рассмотрим разложение определителя матрицы Λ^{-1} по j -ой строке

$$|\Lambda^{-1}| = \sum_{m=1}^p q_{jm} (-1)^{j+m} |(\Lambda^{-1})_{jm}|.$$

Так как параметр q_{jk} отсутствует в матрицах $(\Lambda^{-1})_{jm}$, то производная

$$\frac{\partial |\Lambda^{-1}|}{\partial q_{jk}} = (-1)^{j+k} |(\Lambda^{-1})_{jk}| = \lambda_{jk} |\Lambda|^{-1},$$

где последнее равенство следует из того, что (k, j) -ый элемент обратной к Λ^{-1} матрицы равен $(-1)^{j+k} |(\Lambda^{-1})_{jk}| |\Lambda|$ и совпадает, очевидно, с (k, j) -ым элементом λ_{kj} исходной матрицы Λ . Следовательно, производная функции правдоподобия (с найденными значениями $\mu_j = x_{j\cdot}$) равна

$$\frac{\partial \mathbb{L}}{\partial q_{jk}} = \frac{n}{2} \frac{1}{|\Lambda^{-1}|} \frac{\partial |\Lambda^{-1}|}{\partial q_{jk}} - \frac{1}{2} \sum_{i=1}^n (x_{ji} - x_{j\cdot})(x_{ki} - x_{j\cdot}) = \frac{n}{2} (\lambda_{kj} - l_{kj}).$$

Приравняв к нулю эту производную, получаем доказательство первой части теоремы.

Для доказательства второй части теоремы подставим найденные оценки в выражение (I.20, стр.41). Необходимо показать только, что

$$\sum_{i=1}^n (\vec{x}_i - \vec{x}_{\cdot})' L^{-1} (\vec{x}_i - \vec{x}_{\cdot}) = np.$$

Представим матрицу ковариаций L в виде

$$L = \frac{1}{n} \sum_{i=1}^n (\vec{x}_i - \vec{x}_{\cdot})(\vec{x}_i - \vec{x}_{\cdot})'.$$

Поскольку скалярная величина совпадает со своим следом и внутри функции следа можно переставлять матрицы, то

$$\begin{aligned} \sum_{i=1}^n (\vec{x}_i - \vec{x}_{\cdot})' L^{-1} (\vec{x}_i - \vec{x}_{\cdot}) &= \\ &= \text{tr} \left(\sum_{i=1}^n L^{-1} (\vec{x}_i - \vec{x}_{\cdot})(\vec{x}_i - \vec{x}_{\cdot})' \right) = \text{tr}(n \mathbf{I}) = np. \end{aligned}$$

⊗

Доказательство Лемм I.1, I.2.

Найдем ковариацию $\mathbf{Cov}(l_{12}, l_{11})$ – остальные показатели находятся аналогично. Заметим, что выборочные вторые моменты зависят только от разностей $x_{ji} - x_{j\cdot}$, поэтому заменой $(x_{ji} - \mu_j) \mapsto x_{ji}$ общий случай можно свести к случаю равенства нулю всех средних: $\mu_j = 0, j = \overline{1, p}$.

Сначала вычислим среднее значение

$$\begin{aligned} \mathbf{E} l_{12} &= \mathbf{E} \left[\frac{1}{n} \sum_i X_{1i} X_{2i} - X_{1\cdot} X_{2\cdot} \right] = \\ &= \frac{1}{n} \sum_i \mathbf{E}[X_{1i} X_{2i}] - \frac{1}{n^2} \sum_{jk} \mathbf{E}[X_{1j} X_{2k}]. \end{aligned}$$

В первой сумме здесь все слагаемые равны a_{11} . Во второй сумме слагаемые с разными индексами j, k равны нулю в силу независимости наблюдений в разных экспериментах и предположения равенства нулю средних: $\mathbf{E}(X_{1j} X_{2k}) = (\mathbf{E} X_{1j})(\mathbf{E} X_{2k}) = 0$. Поэтому количество слагаемых равно n и

$$\mathbf{E} l_{12} = a_{11} - \frac{1}{n} a_{11} = \frac{n-1}{n} a_{11} = \frac{n-1}{n} \lambda_{12}.$$

Аналогично $\mathbf{E} l_{11} = \frac{n-1}{n} a_{20} = \frac{n-1}{n} \lambda_{11}$.

Среднее значение произведения

$$\begin{aligned} \mathbf{E}(l_{12} l_{11}) &= \mathbf{E} \left(\frac{1}{n} \sum_i X_{1i} X_{2i} - X_{1\cdot} X_{2\cdot} \right) \left(\frac{1}{n} \sum_l X_{1l}^2 - X_{1\cdot}^2 \right) = \\ &= \mathbf{E} \left(\frac{1}{n} \sum_i X_{1i} X_{2i} - \frac{1}{n^2} \sum_{jk} X_{1j} X_{2k} \right) \cdot \left(\frac{1}{n} \sum_l X_{1l}^2 - \frac{1}{n^2} \sum_{ms} X_{1m} X_{1s} \right) = \\ &= \frac{1}{n^2} \sum_{il} \mathbf{E}[X_{1i} X_{2i} X_{1l}^2] - \frac{1}{n^3} \sum_{jkl} \mathbf{E}[X_{1j} X_{2k} X_{1l}^2] - \\ &\quad - \frac{1}{n^3} \sum_{ims} \mathbf{E}[X_{1i} X_{2i} X_{1m} X_{1s}] + \frac{1}{n^4} \sum_{jkm s} \mathbf{E}[X_{1j} X_{2k} X_{1m} X_{1s}]. \end{aligned}$$

Первая сумма в последнем соотношении содержит n слагаемых $\mathbf{E}[X_{1i} X_{2i} X_{1i}^2] = a_{31}$ с равными индексами $i = l$ и $n^2 - n$ слагаемых $\mathbf{E}[X_{1i} X_{2i} X_{1l}^2] = \mathbf{E}[X_{1i} X_{2i}] \mathbf{E}[X_{1l}^2] = a_{11} a_{20}$ с неравными индексами $i \neq l$. Поэтому эта сумма равна $na_{31} + n(n-1)a_{11}a_{20}$.

Этому же значению равны две следующие суммы. Например, в третьей сумме слагаемые с неравными индексами $m \neq s, m \neq i$ пропадают в силу независимости наблюдений и равенства нулю средних значений (см. вычисления $\mathbf{E} l_{12}$).

В четвертой сумме пропадут все слагаемые, в которых хотя бы один из индексов отличен от трех других. Поэтому в этой сумме не равны нулю слагаемые, у которых индексы удовлетворяют четырем соотношениям:

$$j = k = m = s \text{ (всего } n \text{ слагаемых, равных } a_{31}) ,$$

$$j = k \neq m = s, j = m \neq k = s, j = s \neq m = k$$

(всего $3(n^2 - n)$ слагаемых, равных $a_{11}a_{20}$). Следовательно, четвертая сумма равна $na_{31} + 3n(n-1)a_{11}a_{20}$ или

$$na_{31} + n(n-1)a_{11}a_{20} + 2n(n-1)a_{11}a_{20}.$$

Таким образом,

$$\begin{aligned} \mathbf{E}(l_{12}l_{11}) &= \\ &= \left(1 - \frac{2}{n} + \frac{1}{n^2}\right) \left(\frac{1}{n}a_{31} + \frac{(n-1)}{n}a_{11}a_{20}\right) + 2\frac{n-1}{n^3}a_{11}a_{20} = \\ &= \frac{1}{n} \left(\frac{n-1}{n}\right)^2 a_{31} + \left(\frac{n-1}{n}\right)^3 a_{11}a_{20} + 2\frac{n-1}{n^3}a_{11}a_{20}. \end{aligned}$$

Ковариация

$$\begin{aligned} \mathbf{Cov}(l_{12}, l_{11}) &= \mathbf{E}(l_{12}l_{11}) - \mathbf{E} l_{12} \mathbf{E} l_{11} = \\ &= \mathbf{E}(l_{12}l_{11}) - \left(\frac{n-1}{n}\right)^2 a_{11}a_{20} = \\ &= \frac{1}{n} \left(\frac{n-1}{n}\right)^2 (a_{31} - a_{11}a_{20}) + 2\frac{n-1}{n^3}a_{11}a_{20}, \end{aligned}$$

что завершает доказательство леммы.

⊗

До к а з а т е л ь с т в о Теоремы I.5.

Воспользуемся дельта-методом (Теорема I.3, стр.9). Для этого рассмотрим функцию трех переменных

$$r = \frac{u_3}{\sqrt{u_1 u_2}},$$

которая в точке $\vec{u} = \vec{u}^*$ с координатами

$$u_1^* = \lambda_{11} = \sigma_1^2, \quad u_2^* = \lambda_{22} = \sigma_2^2, \quad u_3^* = \lambda_{12}$$

совпадает с истинным коэффициентом корреляции: $r = \rho$. Производная r по u_1 в точке $\vec{u} = \vec{u}^*$ равна

$$\left. \frac{\partial r}{\partial u_1} \right|_{\vec{u}=\vec{u}^*} = - \frac{1}{2} \frac{u_3}{\sqrt{u_1^3 u_2}} \Big|_{\vec{u}=\vec{u}^*} = - \frac{\rho}{2\sigma_1^2}$$

Аналогично

$$\left. \frac{\partial r}{\partial u_2} \right|_{\vec{u}=\vec{u}^*} = - \frac{\rho}{2\sigma_2^2}, \quad \left. \frac{\partial r}{\partial u_3} \right|_{\vec{u}=\vec{u}^*} = \frac{1}{\sigma_1 \sigma_2}.$$

В соответствии с дельта-методом, коэффициент r асимптотически нормален с дисперсией

$$\tau_\rho^2 = \frac{1}{4} \left(\frac{\rho}{\sigma_1^2}, \frac{\rho}{\sigma_2^2}, -\frac{2}{\sigma_1 \sigma_2} \right) Q \begin{pmatrix} \rho/\sigma_1^2 \\ \rho/\sigma_2^2 \\ -2/\sigma_1 \sigma_2 \end{pmatrix}.$$

Матрица $Q = \mathbf{Cov}(l_{11}, l_{22}, l_{12})$ (см. лемму I.2, стр.7). Доказательство теоремы завершается простыми арифметическими преобразованиями.

⊗

Доказательство Теоремы I.6.

Вычислим вероятность отвержения гипотезы в предположении, что истинный коэффициент корреляции $\rho \leq \rho_0$:

$$\mathbf{P} \left\{ \sqrt{n} \frac{r_n - \rho_0}{\hat{d}} > t^{(\alpha)} \right\} \leq \mathbf{P} \left\{ \sqrt{n} \frac{r_n - \rho}{\hat{d}} > t^{(\alpha)} \right\},$$

так как $(r_n - \rho_0) \leq (r_n - \rho)$. В силу замечания Z7, стр.10, последняя вероятность при $n \rightarrow \infty$ стремится к $1 - \Phi(t^{(\alpha)}) = \alpha$. При $\rho = \rho_0$ асимптотическая вероятность отвержения гипотезы будет очевидно равна α .

⊗

Доказательство Леммы I.3 и Теоремы I.7.

Найдем, например, значение четвертого момента a_{31} . Можно положить средние $\mu_1 = \mu_2 = 0$ и дисперсии $\sigma_1^2 = \sigma_2^2 = 1$. В этом случае характеристическая функция двумерного нормального закона равна

$$\varphi(t_1, t_2) = \exp\left\{-\frac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right\}.$$

Известно, что $(k+j)$ -ые смешанные моменты случайного вектора связаны с $(k+j)$ -ой смешанной производной в нуле его характеристической функции. В частности,

$$a_{31} = \frac{1}{i^4} \frac{\partial^4 \varphi(t_1, t_2)}{\partial t_1^3 \partial t_2} \Big|_{t_1=t_2=0}.$$

Замечая, что $\dot{\varphi}_{t_1} = -(\rho t_2 + t_1)\varphi$, последовательно находим

$$\ddot{\varphi}_{t_1^2} = -\varphi + (\rho t_2 + t_1)^2 \varphi,$$

$$\dot{\ddot{\varphi}}_{t_1^3} = 3(\rho t_2 + t_1)\varphi - (\rho t_2 + t_1)^3 \varphi,$$

$$\ddot{\ddot{\varphi}}_{t_1^4} = 3\rho\varphi - 3(\rho t_2 + t_1)(\rho t_1 + t_2)\varphi - 3\rho(\rho t_2 + t_1)^2 \varphi +$$

$$+(\rho t_2 + t_1)^3(\rho t_1 + t_2)\varphi.$$

При $t_1 = t_2 = 0$ функция $\varphi = 1$, следовательно, $a_{31} = 3\rho$.

Аналогично получаются значения всех остальных моментов. Дальнейшее доказательство вполне очевидно.

⊗

Д о к а з а т е л ь с т в о значимости критерия Фишера (стр.16).

Гипотеза **H** будет отвергаться при $\alpha_{кр} < \alpha$, то есть при

$$Z < \mu_{z0} + \frac{\tau_{z0}}{\sqrt{n-3}} \Phi^{-1}(\alpha),$$

где $\Phi^{-1}(\alpha)$ – α -квантиль нормального закона. Поэтому вероятность отвергнуть гипотезу, когда истинный коэффициент корреляции $\rho > \rho_0$, равна

$$\mathbf{P} \left\{ \frac{z - \mu_{z0}}{\tau_{z0}} \sqrt{n-3} < \Phi^{-1}(\alpha) \middle| \rho \right\} \leq$$

$$\leq \mathbf{P} \left\{ \frac{z - \mu_z}{\tau_z} \sqrt{n-3} < \Phi^{-1}(\alpha) \middle| \rho \right\} \stackrel{n \rightarrow \infty}{\simeq} \Phi(\Phi^{-1}(\alpha)) = \alpha,$$

так как преобразование Фишера монотонно возрастает по ρ .

⊗

Д о к а з а т е л ь с т в о Теоремы I.11.

I. Покажем сначала, что ковариационная матрица вектора $\vec{\theta}_n = (X_{1\cdot}, X_{2\cdot}, l_{22}, l_{12})'$ с точностью до сомножителя $\frac{n-1}{n}$ равна

$$\frac{1}{n} D_{\vartheta} = \frac{1}{n} \begin{pmatrix} a_{20} & a_{11} & a_{12} & a_{21} \\ a_{11} & a_{02} & a_{03} & a_{12} \\ a_{12} & a_{03} & a_{04} - a_{02}^2 & a_{13} - a_{11}a_{02} \\ a_{21} & a_{12} & a_{13} - a_{11}a_{02} & a_{22} - a_{11}^2 \end{pmatrix}.$$

Правая нижняя часть матрицы D_ϑ – это просто матрица Q без первой строки и первого столбца. Поэтому нам осталось вычислить только ковариацию $\mathbf{Cov}(X_{1\cdot}, X_{2\cdot})$ и смешанные ковариации $(X_{1\cdot}, X_{2\cdot})$ и (l_{22}, l_{12}) . Найдем, например, ковариации $\mathbf{Cov}(X_{1\cdot}, X_{2\cdot})$ и $\mathbf{Cov}(X_{1\cdot}, l_{12})$. Так как выборочный центральный момент l_{12} инвариантен относительно сдвига, а ковариации

$$\mathbf{Cov}(X_{1\cdot}, X_{2\cdot}) = \mathbf{E}(X_{1\cdot} - \mu_1)(X_{2\cdot} - \mu_2),$$

$$\mathbf{Cov}(X_{1\cdot}, l_{12}) = \mathbf{E}(X_{1\cdot} - \mu_1)(l_{12} - \mathbf{E}(l_{12})),$$

то можно предположить, что средние значения $\mu_1 = \mu_2 = 0$. Поэтому ковариация

$$\mathbf{Cov}(X_{1\cdot}, X_{2\cdot}) = \mathbf{E}(X_{1\cdot} X_{2\cdot}) = \frac{1}{n^2} \sum_{i,j=1}^n \mathbf{E}(X_{1i} X_{2j}) = \frac{1}{n} a_{11},$$

где в двойной сумме все слагаемые с разными индексами равны нулю в силу независимости разных наблюдений: $\mathbf{E}(X_{1i} X_{2j}) = \mathbf{E}(X_{1i}) \mathbf{E}(X_{2j}) = 0$, а слагаемые с одинаковыми индексами (их всего n штук) равны второму смешанному моменту: $\mathbf{E}(X_{1i} X_{2i}) = a_{11}$.

Ковариация

$$\begin{aligned} \mathbf{Cov}(X_{1\cdot}, l_{12}) &= \mathbf{E}\{X_{1\cdot}(l_{12} - \mathbf{E} l_{12})\} = \mathbf{E}(X_{1\cdot} l_{12}) = \\ &= \mathbf{E} \left\{ \frac{1}{n} \sum_i X_{1i} \left(\frac{1}{n} \sum_j X_{1j} X_{2j} - \frac{1}{n} \sum_k X_{1k} \cdot \frac{1}{n} \sum_l X_{2l} \right) \right\} \\ &= \frac{1}{n^2} \sum_{ij} \mathbf{E}(X_{1i} X_{1j} X_{2j}) - \frac{1}{n^3} \sum_{ikl} \mathbf{E}(X_{1i} X_{1k} X_{2l}) = \\ &= \frac{1}{n} a_{21} - \frac{1}{n^2} a_{21} = \frac{n-1}{n^2} a_{21}, \end{aligned}$$

поскольку снова каждая из сумм содержит только n слагаемых с одинаковыми индексами и эти слагаемые равны моменту $a_{21} = \mathbf{E}(X_{1i}^2 X_{2i})$.

Подставим в матрицу D_ϑ значения смешанных моментов нормально-го распределения. В Лемме I.3 были найдены моменты четвертого порядка. Вторые моменты по определению равны заявленным в теореме значениям:

$$a_{20} = \mathbf{E}(X_1^2) = \sigma_1^2, \quad a_{02} = \mathbf{E}(X_2^2) = \sigma_2^2,$$

$$a_{11} = \mathbf{E}(X_1 X_2) = \lambda_{12} = \rho \sigma_1 \sigma_2.$$

Центральные смешанные моменты третьего порядка нормального случайного вектора, как и все нечетные моменты, равны нулю. Последний факт можно доказать очень простым способом. Любой $(k+m)$ -ый момент случайного вектора (X_1, X_2) с нулевыми средними связан с его характеристической функцией $\varphi(t_1, t_2)$ соотношением

$$a_{km} = \frac{1}{i^{(k+m)}} \left. \frac{\partial^{(k+m)} \varphi}{\partial t_1^k \partial t_2^m} \right|_{t_1=t_2=0}.$$

Характеристическая функция нормального вектора с нулевыми средними значениями не содержит мнимую единицу [10, стр.73, (IV.1)]. Следовательно, при нечётном $(k+m)$ правая часть предыдущего выражения будет чисто мнимой величиной, а левая часть всегда действительна, что возможно только при нулевом значении обеих частей.

II. Снова воспользуемся дельта-методом (Теорема I.3, стр.9). Для этого рассмотрим функцию четырех аргументов

$$h(u_1, u_2, u_3, u_4) = u_1 + \frac{u_4}{u_3}(x - u_2).$$

При

$$u_1^* = \mu_1, \quad u_2^* = \mu_2, \quad u_3^* = \sigma_2^2, \quad u_4^* = \lambda_{12} = \rho\sigma_1\sigma_2$$

значение функции h как раз равно регрессии X_1 на $X_2 = x$. Найдём все её четыре частных производные в точке $(u_1^*, u_2^*, u_3^*, u_4^*)$.

$$\begin{aligned} \dot{h}_{u_1} &= 1, & \dot{h}_{u_3} &= -\frac{u_4}{u_3^2}(x - u_2) = -\frac{\rho\sigma_1}{\sigma_2^2}(x - \mu_2), \\ \dot{h}_{u_2} &= -\frac{u_4}{u_3} = -\frac{\rho\sigma_1}{\sigma_2}, & \dot{h}_{u_4} &= \frac{1}{u_3}(x - u_2) = \frac{1}{\sigma_2^2}(x - \mu_2). \end{aligned}$$

Таким образом, градиент h в точке $(u_1^*, u_2^*, u_3^*, u_4^*)$ равен

$$\vec{\phi} = (1, -\frac{\rho\sigma_1}{\sigma_2}, -\frac{\rho\sigma_1}{\sigma_2^2}(x - \mu_2), \frac{1}{\sigma_2^2}(x - \mu_2)).$$

В соответствии с дельта-методом асимптотическая дисперсия h равна квадратичной форме

$$\begin{aligned} \vec{\phi}' D_{\vartheta} \vec{\phi} &= 1^2 \cdot \sigma_1^2 + \left(-\frac{\rho\sigma_1}{\sigma_2}\right)^2 \cdot \sigma_2^2 + \left(-\frac{\rho\sigma_1}{\sigma_2^2}(x - \mu_2)\right)^2 \cdot 2\sigma_2^4 + \\ &+ \left(\frac{1}{\sigma_2^2}(x - \mu_2)\right)^2 \cdot (\rho^2 + 1)\sigma_1^2\sigma_2^2 + 2 \cdot 1 \cdot \left(-\frac{\rho\sigma_1}{\sigma_2}\right) \cdot \rho\sigma_1\sigma_2 + \\ &+ 2 \cdot \left(-\frac{\rho\sigma_1}{\sigma_2^2}(x - \mu_2)\right) \cdot \left(\frac{1}{\sigma_2^2}(x - \mu_2)\right) \cdot 2\rho\sigma_1\sigma_2^3. \end{aligned}$$

Приведя подобные члены, получаем

$$\vec{\phi}' D_{\vartheta} \vec{\phi} = \sigma_1^2 (1 - \rho^2) \left(1 + \frac{(x - \mu_2)^2}{\sigma_2^2} \right),$$

что (после замены истинных значений параметров на их выборочные аналоги) совпадает с выражением (I.8) для асимптотической дисперсии линейной регрессии.

⊗

Доказательство Теоремы I.13.

Найдем характеристическую функцию вектора рассматриваемых статистик. Пусть $\vec{x}_i = (x_{1i}, \dots, x_{pi})'$ – выборочные данные, полученные в i -ом эксперименте. Аргумент характеристической функции, относящийся к выборочному среднему j -ой компоненты $x_{j\cdot}$, обозначим a_j , а аргумент, соответствующий ковариации l_{jk} , обозначим t_{jk} . Тогда характеристическая функция (в предположении, что вектор истинных средних $\vec{\mu} = \vec{0}$) равна

$$\begin{aligned} \varphi(a_1, \dots, a_p, t_{11}, t_{12}, \dots, t_{pp}) &= \left(\sqrt{(2\pi)^p |\Lambda|} \right)^{-n} \cdot \\ &\cdot \int_{\mathcal{R}^{np}} \exp \left\{ i(\vec{a}' \vec{x}_{\cdot} + \text{tr } TL) - \frac{1}{2} \sum_{i=1}^n \vec{x}_i' \Lambda^{-1} \vec{x}_i \right\} d\vec{x}_1 \dots d\vec{x}_n. \end{aligned}$$

Здесь выражение $\text{tr } TL$ с симметричной матрицей

$$T = \begin{pmatrix} t_{11} & \frac{1}{2}t_{12} & \dots & \frac{1}{2}t_{1k} \\ \frac{1}{2}t_{12} & t_{11} & \dots & \frac{1}{2}t_{2k} \\ \dots & \dots & \dots & \dots \\ \frac{1}{2}t_{1k} & \frac{1}{2}t_{1k} & \dots & t_{kk} \end{pmatrix}$$

представляет собой более короткую запись суммы $\sum_{j \leq k} t_{jk} l_{jk}$.

Перейдем сначала в этом интеграле к новым переменным, производя поворот к главным осям квадратичной формы. Для положительно определенной матрицы Λ^{-1} найдется ортогональная матрица Q такая, что $Q' \Lambda^{-1} Q = \Gamma = \text{diag}(\gamma_1^2, \dots, \gamma_p^2)$. Для каждого $i = \overline{1, n}$ положим

$$\vec{x}_i = Q \Gamma^{-\frac{1}{2}} \vec{y}_i$$

с матрицей $\Gamma^{-\frac{1}{2}} = \text{diag}(\frac{1}{\gamma_1}, \dots, \frac{1}{\gamma_p})$. Якобиан такой подстановки равен n -ой степени (так как замена производится для всех $i = \overline{1, n}$) определителя

матрицы преобразования $|Q\Gamma^{-\frac{1}{2}}| = |\Gamma|^{-\frac{1}{2}}$. Из определения матрицы Q видно, что $|\Gamma| = |\Lambda^{-1}|$. Следовательно, якобиан равен $(\sqrt{|\Lambda|})^n$.

Квадратичная форма под знаком экспоненты при таком преобразовании переходит в сумму квадратов компонент:

$$\vec{x}'_i \Lambda^{-1} \vec{x}_i = \vec{y}'_i \Gamma^{-\frac{1}{2}} Q' \Lambda^{-1} Q \Gamma^{-\frac{1}{2}} \vec{y}_i = \vec{y}'_i \vec{y}_i,$$

что позволит нам в дальнейшем разделить переменные интегрирования.

Вектор средних представляется в виде

$$\vec{x}_\bullet = \frac{1}{n} \sum_i \vec{x}_i = Q \Gamma^{-\frac{1}{2}} \frac{1}{n} \sum_i \vec{y}_i = Q \Gamma^{-\frac{1}{2}} \vec{y}_\bullet.$$

Матрица ковариаций

$$\begin{aligned} L &= \frac{1}{n} \sum_i \vec{x}_i \vec{x}'_i - \vec{x}_\bullet \vec{x}_\bullet' = Q \Gamma^{-\frac{1}{2}} \left(\frac{1}{n} \sum_i \vec{y}_i \vec{y}'_i - \vec{y}_\bullet \vec{y}_\bullet' \right) \Gamma^{-\frac{1}{2}} Q' \\ &= Q \Gamma^{-\frac{1}{2}} L \Gamma^{-\frac{1}{2}} Q', \end{aligned}$$

где для матрицы ковариаций относительно переменных y оставлено тоже обозначение L . В силу перестановочности матриц внутри следа

$$\text{tr} TL = \text{tr}(T Q \Gamma^{-\frac{1}{2}} L \Gamma^{-\frac{1}{2}} Q') = \text{tr}(\Gamma^{-\frac{1}{2}} Q' T Q \Gamma^{-\frac{1}{2}} L).$$

Теперь, если ввести новый вектор $\vec{\alpha}' = \vec{a}' Q \Gamma^{-\frac{1}{2}}$ и новую матрицу $\mathfrak{S} = \Gamma^{-\frac{1}{2}} Q' T Q \Gamma^{-\frac{1}{2}}$, то интеграл, определяющий искомую характеристическую функцию, можно записать в виде

$$\int_{\mathcal{R}^{np}} \exp \left\{ \mathbf{i}(\vec{\alpha}' \vec{y}_\bullet + \text{tr} \mathfrak{S} L) - \frac{1}{2} \sum_{i=1}^n \vec{y}'_i \vec{y}_i \right\} d\vec{y}_1 \dots d\vec{y}_n.$$

Следующим шагом наших вычислений будет переход к таким новым переменным, при которых p -мерный вектор $\vec{y}_\bullet$, зависящий от всей совокупности переменных интегрирования, перейдет в p -мерный вектор $\vec{z}_1$, зависящий только от первых p компонент. Для этого рассмотрим n -мерный вектор $\vec{Y}_j = (y_{j1}, \dots, y_{jn})'$, содержащий n измерений по j -ому признаку (в отличие от вектора $\vec{y}_i$, содержащего преобразованные измерения всех p признаков в i -ом эксперименте). Произведем линейное преобразование

$\vec{Z}_j = U \vec{Y}_j$ с ортогональной матрицей $U = U^{n \cdot n}$, у которой первая строка состоит из чисел $1/\sqrt{n}$:

$$\vec{Z}_j = \begin{pmatrix} z_{j1} \\ \vdots \\ z_{jn} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \end{pmatrix} \begin{pmatrix} y_{j1} \\ \vdots \\ y_{jn} \end{pmatrix}.$$

Таким образом, $z_{j1} = 1/\sqrt{n} \sum_i y_{ji} = \sqrt{n} y_{j\bullet}$.

Смешанный момент (в силу ортогональности U)

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n y_{ji} y_{ki} &= \frac{1}{n} \vec{Y}_j' \vec{Y}_k = \frac{1}{n} \vec{Z}_j' U' U \vec{Z}_k = \frac{1}{n} \vec{Z}_j' \vec{Z}_k = \\ &= \frac{1}{n} \sum_{i=1}^n z_{ji} z_{ki}. \end{aligned}$$

В частности, сумма квадратов

$$\sum_{i=1}^n \vec{y}_i' \vec{y}_i = \sum_{j=1}^p \sum_{i=1}^n y_{ji} y_{ji} = \sum_{j=1}^p \sum_{i=1}^n z_{ji} z_{ji} = \sum_{i=1}^n \vec{z}_i' \vec{z}_i,$$

где p -мерный вектор $\vec{z}_i = (z_{1i}, \dots, z_{pi})'$.

Ковариация преобразуется следующим образом:

$$l_{jk} = \frac{1}{n} \sum_{i=1}^n y_{ji} y_{ki} - y_{j\bullet} y_{k\bullet} = \frac{1}{n} \sum_{i=1}^n z_{ji} z_{ki} - \frac{1}{n} z_{j1} z_{k1} = \frac{1}{n} \sum_{i=2}^n z_{ji} z_{ki}.$$

А матрица ковариаций равна

$$L = \frac{1}{n} \sum_{i=1}^n \vec{y}_i \vec{y}_i' - \vec{y}_{\bullet} \vec{y}_{\bullet}' = \frac{1}{n} \sum_{i=2}^n \vec{z}_i \vec{z}_i'$$

и, таким образом, не содержит компонент вектора $\vec{z}_1$.

Поскольку якобиан ортогонального преобразования равен 1, то нам удалось представить искомый интеграл в виде произведения двух интегралов. Первый из этих интегралов (по компонентам p -мерного вектора $\vec{z}_1$) равен

$$J_1 = \int_{\mathcal{R}^p} \exp \left\{ \mathbf{i} \frac{1}{\sqrt{n}} \vec{\alpha}' \vec{z}_1 - \frac{1}{2} \vec{z}_1' \vec{z}_1 \right\} d\vec{z}_1$$

и по определению представляет собой умноженную на $\sqrt{(2\pi)^p}$ характеристическую функцию p -мерного нормального вектора $\mathcal{N}_p(\vec{0}, \mathbf{I})$, вычисленную в точке $\vec{\alpha}/\sqrt{n}$. Поэтому (см. [10, стр.73, Теорема IV.3])

$$\frac{1}{\sqrt{(2\pi)^p}} J_1 = \exp \left\{ -\frac{1}{2n} \vec{\alpha}' \vec{\alpha} \right\}.$$

Произведя обратную замену $\vec{\alpha}' = \vec{a}' Q \Gamma^{-\frac{1}{2}}$ и замечая, что $\Lambda = Q \Gamma^{-1} Q' = Q \Gamma^{-\frac{1}{2}} \Gamma^{-\frac{1}{2}} Q'$, окончательно получаем, что

$$\frac{1}{\sqrt{(2\pi)^p}} J_1 = \exp \left\{ -\frac{1}{2} \vec{a}' \left(\frac{1}{n} \Lambda \right) \vec{a} \right\},$$

то есть совпадает с характеристической функцией нормального $\mathcal{N}_p(\vec{0}, \frac{1}{n} \Lambda)$ распределения.

Итак, искомая характеристическая функция φ представляется в виде произведения двух функций $\varphi_1 \cdot \varphi_2$, одна из которых φ_1 зависит только от переменных $a_1, \dots, a_p$, соответствующих компонентам вектора средних, а вторая φ_2 зависит только от переменных T , соответствующих компонентам вектора ковариаций $\vec{\mathcal{L}}$. Поскольку функция φ_1 есть характеристическая функция нормального $(\vec{0}, \frac{1}{n} \Lambda)$ закона, то и φ_2 тоже характеристическая функция, а следовательно, первые два пункта теоремы нами доказаны.

Для доказательства третьего пункта упростим запись интеграла, входящего в определение φ_2 :

$$\begin{aligned} J_2 &= \int_{\mathcal{R}^{p(n-1)}} \exp \left\{ \mathbf{i} \frac{1}{n} \sum_{i=2}^n \text{tr}(\Im \vec{z}_i \vec{z}_i') - \frac{1}{2} \sum_{i=2}^n \vec{z}_i' \vec{z}_i \right\} d\vec{z}_2 \dots d\vec{z}_n \\ &= \prod_{i=2}^n \int_{\mathcal{R}^p} \exp \left\{ \mathbf{i} \frac{1}{n} \text{tr}(\Im \vec{z}_i \vec{z}_i') - \frac{1}{2} \vec{z}_i' \vec{z}_i \right\} d\vec{z}_i = \\ &= \left(\int_{\mathcal{R}^p} \exp \left\{ \mathbf{i} \frac{1}{n} \text{tr}(\Im \vec{z} \vec{z}') - \frac{1}{2} \vec{z}' \vec{z} \right\} d\vec{z} \right)^{n-1}. \end{aligned}$$

Так как

$$\text{tr}(\Im \vec{z} \vec{z}') = \text{tr}(\vec{z}' \Im \vec{z}) = \vec{z}' \Im \vec{z} = \vec{z}' \Gamma^{-\frac{1}{2}} Q' T Q \Gamma^{-\frac{1}{2}} \vec{z},$$

то после замены $\vec{w} = Q \Gamma^{-\frac{1}{2}} \vec{z}$ (обратное преобразование по отношению к переходу от переменных x к переменным y) приходим к вычислению ин-

теграла

$$\frac{1}{\sqrt{|\Lambda|}} \int_{\mathcal{R}^p} \exp \left\{ \mathbf{i} \frac{1}{n} \vec{w}' T \vec{w} - \frac{1}{2} \vec{w}' \Lambda^{-1} \vec{w} \right\} d\vec{w},$$

который в силу интегрального соотношения (B1)[10, стр.95] с матрицей $A = \Lambda^{-1} - \mathbf{i}2T/n$ равен $(\sqrt{2\pi})^p / \sqrt{|A|}$.

Таким образом, функция φ_2 равна

$$\varphi_2(t_{11}, t_{12}, \dots, t_{pp}) = \frac{\left| \frac{n}{2} \Lambda^{-1} \right|^{\frac{n-1}{2}}}{\left| \frac{n}{2} \Lambda^{-1} - \mathbf{i}T \right|^{\frac{n-1}{2}}}.$$

Для завершения доказательства теоремы покажем, что характеристическая функция плотности, заявленной в третьем пункте, совпадает с φ_2 . Примем без доказательства, что функция $f_n(\vec{\mathcal{L}})$ есть плотность некоторого распределения и соотношение $\int f_n(\vec{\mathcal{L}}) d\vec{\mathcal{L}} = 1$ выполняется для любой положительно определенной матрицы Λ , даже содержащей мнимые компоненты. Таким образом, для любой матрицы B ($= \Lambda^{-1}$) имеет место равенство

$$\int_{L>0} C'_{pn} |L|^{\frac{n-p-2}{2}} \exp \left\{ -\frac{n}{2} \text{tr}(LB) \right\} d\vec{\mathcal{L}} = \frac{1}{|B|^{\frac{n-1}{2}}}$$

с константой $C'_{pn} = C_{pn} |\Lambda|^{\frac{n-1}{2}}$, не зависящей от B . Положив в этом тождестве $B = \Lambda^{-1} - \mathbf{i}2T/n$ и разделив его на $|\Lambda|^{\frac{n-1}{2}}$, получим

$$\int_{L>0} \exp \{ \mathbf{i} \text{tr}(TL) \} C_{pn} |L|^{\frac{n-p-2}{2}} \exp \left\{ -\frac{n}{2} \text{tr}(L\Lambda^{-1}) \right\} d\vec{\mathcal{L}} = \frac{|\Lambda|^{-\frac{n-1}{2}}}{|\Lambda^{-1} - \mathbf{i}\frac{2}{n}T|^{\frac{n-1}{2}}}.$$

Левая часть этого соотношения по определению есть характеристическая функция плотности $f_n(\vec{\mathcal{L}})$, а правая часть совпадает с характеристической функцией φ_2 .

⊗

Доказательство Теоремы I.14.

Имеем для $U = nL$

$$\mathbf{E} |L|^m = \frac{1}{n^{p \cdot m}} \mathbf{E} |U|^m = \frac{C_{pn}}{n^{p \cdot m}} \int_{U>0} |U|^m |U|^{\frac{n-p-2}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(U\Lambda^{-1}) \right\} dU,$$

где dU — символическая запись для выражения $dl_{11} \dots dl_{pp}$. Поскольку при $m = 0$ левая часть равна единице, то отсюда получаем тождество

$$\int_{W>0} |U|^{\frac{n-p-2}{2}} \exp \left\{ -\frac{1}{2} \text{tr} (U \Lambda^{-1}) \right\} dU = (2^p |\Lambda|)^{\frac{n-1}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^p \Gamma\left(\frac{n-j}{2}\right),$$

справедливое при все значениях n . При $n = n + 2m$ интеграл в левой части этого соотношения полностью совпадает с интегралом, определяющим математическое ожидание $\mathbf{E} |L|^m$. Следовательно,

$$\begin{aligned} \mathbf{E} |L|^m &= \frac{C_{pn}}{n^{\frac{p(n-1)}{2} + mp}} (2^p |\Lambda|)^{\frac{n-1}{2} + m} \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^p \Gamma\left(\frac{n-j}{2} + m\right) \\ &= \left(\frac{2^p |\Lambda|}{n^p}\right)^m \prod_{j=1}^p \frac{\Gamma(\frac{n-j}{2} + m)}{\Gamma(\frac{n-j}{2})}. \end{aligned}$$

Для того, чтобы отсюда найти выражение для среднего значения, заметим, что $\Gamma(\frac{n-j}{2} + 1) = \frac{(n-j)}{2} \Gamma(\frac{n-j}{2})$. Поэтому

$$\mathbf{E} |L| = \frac{2^p |\Lambda|}{n^p} \prod_{j=1}^p \frac{(n-j)}{2} = |\Lambda| \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{p}{n}\right).$$

С другой стороны, так как $\Gamma\left(\frac{n-j}{2} + 2\right) = \left(\frac{n-j}{2} + 1\right) \frac{(n-j)}{2} \Gamma\left(\frac{n-j}{2}\right)$, то

$$\mathbf{E} |L|^2 = \frac{2^{2p} |\Lambda|^2}{n^{2p}} \prod_{j=1}^p \left(\frac{n-j}{2} + 1\right) \frac{(n-j)}{2} = |\Lambda|^2 \prod_{j=1}^p \left(1 - \frac{j-2}{n}\right) \left(1 - \frac{j}{n}\right).$$

Следовательно, дисперсия

$$\begin{aligned} \mathbf{D} |L| &= \mathbf{E} |L|^2 - (\mathbf{E} |L|)^2 = \\ &= |\Lambda|^2 \prod_{j=1}^p \left(1 - \frac{j}{n}\right) \prod_{k=1}^{p-2} \left(1 - \frac{k}{n}\right) \cdot \left[\left(1 + \frac{1}{n}\right) - \left(1 - \frac{p-1}{n}\right) \left(1 - \frac{p}{n}\right) \right] = \\ &= |\Lambda|^2 \frac{2p}{n} \left[\prod_{j=1}^{p-2} \left(1 - \frac{j}{n}\right) \right]^2 \cdot \left(1 - \frac{p-1}{n}\right) \left(1 - \frac{p}{n}\right) \left(1 - \frac{p-1}{2n}\right). \end{aligned}$$

⊗

Доказательство Теоремы I.15.

Заметим сначала, что так как коэффициент корреляции зависит от вторых выборочных моментов l_{11}, l_{22}, l_{12} , то естественно выразить распределение r через функцию распределения этих моментов:

$$\begin{aligned} \mathbf{P}\{r < x\} &= \mathbf{P}\{l_{11} < \infty, l_{22} < \infty, l_{12} < x\sqrt{l_{11}l_{22}}\} \\ &= \int_0^\infty dl_{11} \int_0^\infty dl_{22} \int_{-\sqrt{l_{11}l_{22}}}^{x\sqrt{l_{11}l_{22}}} f_n(l_{11}, l_{22}, l_{12}) dl_{12} . \end{aligned}$$

Область интегрирования в этом представлении взята в указанном виде по причине того, что функция плотности $f_n > 0$ только при $L > 0$, то есть при $l_{11} \in [0, \infty), l_{22} \in [0, \infty), l_{12} \in (-\sqrt{l_{11}l_{22}}, \sqrt{l_{11}l_{22}})$. Для получения функции плотности r продифференцируем последнее выражение по x . Таким образом, плотность распределения r (после замены аргумента x на r) равна

$$f(r) = \int_0^\infty dl_{11} \int_0^\infty dl_{22} \sqrt{l_{11}l_{22}} f_n(l_{11}, l_{22}, r\sqrt{l_{11}l_{22}}) . \quad (\text{I.21})$$

Выпишем функцию плотности L в частном случае $p = 2$. Так как определитель $|L| = l_{11}l_{22}(1 - r^2)$, а след

$$\text{tr}(L\Lambda^{-1}) = -\frac{1}{(1 - \rho^2)} \left(\frac{l_{11}}{\sigma_1^2} - 2\rho \frac{l_{12}}{\sigma_1\sigma_2} + \frac{l_{22}}{\sigma_2^2} \right) ,$$

то функция плотности L (при указанных выше ограничениях на свои аргументы) равна

$$\begin{aligned} f_n(l_{11}, l_{22}, l_{12}) &= C_n (l_{11} l_{22} (1 - r^2))^{\frac{n-4}{2}} \cdot \\ &\exp \left\{ -\frac{n}{2(1 - \rho^2)} \left(\frac{l_{11}}{\sigma_1^2} - 2\rho \frac{l_{12}}{\sigma_1\sigma_2} + \frac{l_{22}}{\sigma_2^2} \right) \right\} , \\ \text{где} \quad C_n &= \frac{1}{\sqrt{\pi}} \left(\frac{n}{2\sigma_1\sigma_2\sqrt{1 - \rho^2}} \right)^{n-1} \frac{1}{\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})} . \end{aligned}$$

Подставляя сюда $l_{12} = r\sqrt{l_{11}l_{22}}$, как это необходимо для вычисления (I.21),

получаем подынтегральное выражение (I.21)

$$\sqrt{l_{11}l_{22}} f_n(l_{11}, l_{22}, r\sqrt{l_{11}l_{22}}) = C_n (l_{11} l_{22})^{\frac{n-3}{2}} (1-r^2)^{\frac{n-4}{2}} \cdot \\ \exp \left\{ -\frac{n}{2(1-\rho^2)} \left(\frac{l_{11}}{\sigma_1^2} - 2r\rho\sqrt{\frac{l_{11}}{\sigma_1^2}}\sqrt{\frac{l_{22}}{\sigma_2^2}} + \frac{l_{22}}{\sigma_2^2} \right) \right\}.$$

Для дальнейших вычислений удобно произвести замену в интеграле (I.21), положив $u_i = nl_{ii}/(2\sigma_i^2(1-\rho^2))$, $i = 1, 2$. После несложных преобразований получаем, что плотность (I.21) равна

$$f(r) = A \cdot \int_0^\infty du_1 \int_0^\infty du_2 u_1^{(n-3)/2} u_2^{(n-3)/2} e^{-u_1} e^{-u_2} e^{2r\rho u_1 u_2},$$

где сомножитель
$$A = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{\sqrt{\pi}\Gamma(\frac{n-1}{2})\Gamma(\frac{n-2}{2})} (1-r^2)^{\frac{n-4}{2}}.$$

Разложим последний сомножитель подынтегрального выражения в ряд Тейлора:

$$e^{2r\rho u_1 u_2} = \sum_{j=0}^{\infty} \frac{(2r\rho)^j}{j!} u_1^j u_2^j.$$

Для окончательного доказательства теоремы осталось вычислить произведение двух интегралов вида

$$\int_0^\infty u^{(n-3)/2} u^j e^{-u} du,$$

которые есть не что иное, как гамма-функция $\Gamma(\frac{n+j-1}{2})$.

⊗

Доказательство Теоремы I.18.

Воспользуемся стандартной схемой:

а) распределение $\rightarrow$ б) производная $\rightarrow$ в) плотность.

Простым дифференцированием легко устанавливается, что статистика Стьюдента есть возрастающая функция r . Поэтому неравенство $T < t$ эквивалентно неравенству $r < r(t)$, где $r(t) = t/\sqrt{n-2+t^2}$ – решение уравнения $T = t$. Таким образом, функция распределения T равна

$$F_T(t) = \mathbf{P}\{T < t\} = \mathbf{P}\{r < r(t)\} = \int_{-1}^{r(t)} f(r) dr.$$

Производная этой функции равна

$$\begin{aligned}
f_T(t) &= \dot{r}(t) f(r(t)) = \\
&= \frac{n-2}{(n-2+t^2)^{3/2}} \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi}\Gamma(\frac{n-2}{2})} \left[\frac{n-2}{(n-2+t^2)} \right]^{\frac{n-4}{2}} = \\
&= \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi}(n-2)\Gamma(\frac{n-2}{2})} \left(1 + \frac{t^2}{n-2} \right)^{-\frac{n-2+1}{2}},
\end{aligned}$$

что совпадает с функцией плотности распределения Стюдента с $(n-2)$ -мя степенями свободы (см., например, [6, стр.264]).

⊗

Доказательство Теоремы I.16.

Без доказательства примем, что функция распределения $F_n(r|\rho)$ непрерывна по обоим параметрам, возрастает по r и убывает по ρ . Пусть $\underline{\rho} = \underline{\rho}(r)$ – единственное решение по параметру ρ , а $\underline{r} = \underline{r}(\rho)$ – единственное решение по параметру r уравнения

$$F_n(r|\rho) = 1 - \alpha.$$

Поскольку $F_n(r|\rho)$ убывает по ρ , то неравенство $\rho > \underline{\rho}(r)$ эквивалентно

$$F_n(r|\rho) < F_n(r|\underline{\rho}) \quad (= 1 - \alpha = F_n(\underline{r}|\rho)).$$

С другой стороны, так как функция $F_n(r|\rho)$ возрастает по r , то последнее эквивалентно неравенству

$$r < \underline{r}(\rho).$$

Следовательно, так как вероятность

$$\mathbf{P}\{r < \underline{r}(\rho)|\rho\} = F_n(\underline{r}(\rho)|\rho) = 1 - \alpha,$$

то и вероятность

$$\mathbf{P}\{\underline{\rho} < \rho|\rho\} = \mathbf{P}\{r < \underline{r}(\rho)|\rho\} = 1 - \alpha.$$

Утверждение для верхней границы доказывается аналогично.

⊗

Доказательство Теоремы I.21.

Так как квадрат множественной корреляции равен $r_{1*}^2 = 1 - \frac{SS_e}{n s_1^2}$, то

$$\frac{r_{1*}^2}{1 - r_{1*}^2} = \frac{n s_1^2 - SS_e}{SS_e}.$$

Если зафиксировать значения переменных $X_2 = x_2, \dots, X_p = x_p$, то переменная X_1 будет иметь нормальное распределение со средним, равным регрессии X_1 на $X_2 = x_2, \dots, X_p = x_p$, и дисперсией, равной остаточной дисперсии σ_{1*}^2 , не зависящей от $x_2, \dots, x_p$ (см. [10, стр.78, Теорема IV.9]). Если верны предположения гипотезы, то все коэффициенты этой регрессии равны нулю, то есть условное распределение X_1 не зависит от фиксированных значений $x_2, \dots, x_p$. Другими словами, при фиксированных значениях предикторных переменных мы находимся в рамках задачи проверки гипотезы о равенстве нулю значений всех $(p - 1)$ коэффициентов регрессии в модели с детерминированными предикторами. Поэтому условная вероятность (см. [9, стр.19, Теорема II.4])

$$\begin{aligned} \mathbf{P} \left\{ \frac{n-p}{p-1} \frac{r_{1*}^2}{1-r_{1*}^2} > z \mid X_{ji} = x_{ji}, i = \overline{1, n}, j = \overline{2, p} \right\} = \\ = 1 - \mathcal{F}_{p-1, n-p}(z), \end{aligned}$$

при любых значениях условия. Следовательно, этому же значению будет равна и безусловная вероятность.

Так как функция

$$\frac{t^2}{(1-t^2)} = -1 - \frac{1}{(1-t^2)}$$

монотонно возрастает по t^2 (< 1), то при

$$z = \frac{n-p}{p-1} \frac{\hat{r}_{1*}^2}{1-\hat{r}_{1*}^2}$$

найденная вероятность будет совпадать с искомой вероятностью $\mathbf{P}\{\hat{r}_{1*}^2 > \hat{r}_{1*}^2\}$.

⊗

Глава II. РАНГОВАЯ КОРРЕЛЯЦИЯ

Не ранжируй, да не ранжируем
будешь!

Очень часто на практике встречаются ситуации, когда наблюдаемые в эксперименте характеристики упорядочены по возрастанию, т.е. представляют собой результат ранжирования объектов по некоторым признакам (например, по месту, занятому в рейтинге). Кроме того, выборочный коэффициент корреляции весьма чувствителен к наличию в выборке резко выпадающих наблюдений. Поэтому зачастую для того, чтобы устранить это влияние, а также при сильных подозрениях против нормальности распределения выборки осуществляют переход от исходных „непрерывных“ данных к рангам (занятым местам) по каждой из характеристик. Для данных такого типа предложено несколько мер взаимосвязи. Мы рассмотрим две из них – коэффициенты корреляции Спирмена и Кендэлла.

§ 1. Коэффициент корреляции Спирмена

Пусть в эксперименте наблюдались два ряда связанных между собой чисел $(x_1, y_1), \dots, (x_n, y_n)$, где x_i – ранг (место) i -ого объекта по первому признаку, y_i – соответствующий ранг по второму признаку (например, команда „Real“ заняла четвертое место в чемпионате Испании ($x_1 = 4$), при этом её игроки получили самую высокую зарплату ($y_1 = 1$)). Очевидно, что в каждом ряду данных встречаются все числа от 1 до n . Кроме того, заметим, что номер каждому объекту присваивается произвольно, без какой-либо связи с рангами по признакам. Однако всегда удобнее, чтобы по

одному из признаков (например, по X -ам) данные располагались в порядке возрастания.

Определение 1. Коэффициент корреляции Пирсона (см. §1, стр.5) между порядковыми (ранжированными) данными называется коэффициентом корреляции Спирмена.

В следующей теореме приводится более удобный способ вычисления этого коэффициента.

Теорема II.1. Если $d_i = x_i - y_i$ — разность рангов i -ого объекта, то коэффициент корреляции Спирмена равен

$$r_S = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n d_i^2.$$

Из определения коэффициента Спирмена как коэффициента корреляции Пирсона легко вытекают его свойства.

Свойства r_S :

- 1) $-1 \leq r_S \leq 1$;
- 2) $r_S = 1$, только если ранги обоих признаков совпадают;
- 3) $r_S = -1$, только если ранги признаков противоположны.

При малых объемах выборок n можно найти точное распределение r_S . Для этого достаточно перебрать все возможные перестановки рангов и подсчитать соответствующие им вероятности. Например, если наблюдалось три объекта, то имеется всего $6 = 3!$ возможных вариантов расположения мест признака Y . В предположении независимости признаков вероятность каждого из вариантов равна $1/6$. Поэтому распределение коэффициента корреляции Спирмена равно

r_S	-1	-0.5	0.5	1
P	$1/6$	$1/3$	$1/3$	$1/6$
Y	$(3, 2, 1)$	$(2, 3, 1)$ $(3, 1, 2)$	$(1, 3, 2)$ $(2, 1, 3)$	$(1, 2, 3)$

При больших значениях n применяют обычно асимптотическое распределение. Основным интерес здесь представляет проверка гипотезы независимости.

Теорема II.2. Пусть r_S – коэффициент корреляции Спирмена, построенный по выборке объема n из генеральной совокупности с независимыми компонентами. Тогда

$$r_S \rightsquigarrow \mathcal{N}\left(0, \frac{1}{n-1}\right), \quad n \rightarrow \infty.$$

Проверка независимости по коэффициенту Спирмена

Из приведенной теоремы легко следует, что если $\hat{r}_S$ – выборочное значение коэффициента корреляции Спирмена, то при проверке гипотезы независимости признаков критический уровень значимости

$$\alpha_{\text{кр}} = \mathbf{P}\{|r_S| > |\hat{r}_S|\} \approx 2 \left[1 - \Phi\left(\sqrt{n-1} \cdot |\hat{r}_S|\right)\right].$$

Как уже отмечалось, зачастую приходится переходить от исходных „непрерывных“ данных к ранжированным. Поэтому представляет интерес связь рангового коэффициента корреляции с исходным полным коэффициентом корреляции. Без доказательства приведем следующее утверждение.

Теорема II.3. Пусть r_S – ранговый коэффициент корреляции Спирмена, построенный по выборке объема n из двумерной нормальной генеральной совокупности с истинной корреляцией ρ . Тогда при $n \rightarrow \infty$ по вероятности

$$r_S \rightarrow \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right). \quad (\text{II.1})$$

§ 2. Коэффициент корреляции Кендэлла

Для вычисления рангового коэффициента корреляции Кендэлла расположим данные по одному из признаков в порядке возрастания. Для второго признака вычислим количество инверсий¹ Π в ряду данных $y_1, \dots, y_n$.

О п р е д е л е н и е 2. Коэффициент

$$r_K = 1 - \frac{4\Pi}{n(n-1)}$$

называется коэффициентом корреляции Кендэлла.

Свойства этого коэффициента аналогичны свойствам коэффициента Спирмена (см. стр.60). Кроме того, имеет место

Теорема II.4. 1) Пусть r_K – ранговый коэффициент корреляции Кендэлла, построенный по выборке объема n из генеральной совокупности с независимыми компонентами. Тогда

$$r_K \rightsquigarrow \mathcal{N}\left(0, \frac{2(2n+5)}{9n(n-1)}\right).$$

2) Пусть r_K – ранговый коэффициент корреляции Кендэлла, построенный по выборке объема n из нормальной генеральной совокупности с истинным коэффициентом корреляции ρ . Тогда при $n \rightarrow \infty$ по вероятности

$$r_K \rightarrow \frac{2}{\pi} \arcsin(\rho). \quad (\text{II.2})$$

§ 1. Одним из несомненных преимуществ рангового коэффициента корреляции Кендэлла является удобство его вычисления без применения вычислительной техники. Однако коэффициент Спирмена более близок к истинному коэффициенту корреляции ρ . Максимальное отклонение правой

¹Инверсией называется нарушение порядка расположения чисел. Например, в ряду (4,2,1,3) всего 4 инверсии (4,2),(4,3),(4,1),(2,1).

части формулы (II.1) от ρ составляет всего 0.0181, в то время как для соотношения (II.2) это отклонение равно 0.2105.

§ 3. Меры связи для нескольких признаков.

Коэффициент конкордации

При изучении связей трех и более ранжированных признаков применяют коэффициент, который показывает степень отклонения суммы рангов по каждому наблюдению от среднего значения всех рангов. Пусть

- R_{ij} — ранг i -го объекта по j -му признаку,
- $R_i = \sum_{j=1}^p R_{ij}$ — сумма рангов i -го объекта по всем p признакам,
- $\bar{R} = \frac{1}{n} \sum_{i=1}^n R_i = \frac{p(n+1)}{2}$ — среднее значение всех рангов,
- $R_W = \sum_{i=1}^n (R_i - \bar{R})^2$ — квадратическое отклонение рангов объектов от их общей середины.

О п р е д е л е н и е 3. Коэффициент конкордации Кендэлла равен

$$W = \frac{12 R_W}{n(n^2 - 1) p^2},$$

З 2. Если число признаков $p = 2$, то коэффициент конкордации линейно связан с коэффициентом корреляции Спирмена:

$$W = \frac{1}{2}(r_S + 1).$$

Поэтому статистические выводы о связях между двумя показателями, основанные на значениях W и r_S , будут полностью идентичны.

Последний факт подтверждает также приводимая ниже теорема об асимптотическом распределении статистики W .

Теорема II.5. Если все признаки независимы в совокупности, то

$$1) \mathbf{E} W = \frac{1}{p},$$

$$2) W \rightsquigarrow \mathcal{N}\left(\frac{1}{p}, \frac{1}{p^2(n-1)}\right), \quad n \rightarrow \infty.$$

3) Если $\hat{W}$ – выборочное значение коэффициента конкордации, то при проверке гипотезы независимости всех признаков критический уровень значимости равен

$$\begin{aligned} \alpha_{кр} &= \mathbf{P} \left\{ \left| W - \frac{1}{p} \right| > \left| \hat{W} - \frac{1}{p} \right| \right\} \approx \\ &\approx 2 \left[1 - \Phi \left(p\sqrt{n-1} \cdot \left| \hat{W} - \frac{1}{p} \right| \right) \right]. \end{aligned}$$

§ 3. Равенство нулю коэффициента корреляции не означает независимости наблюдаемых показателей. Поэтому с помощью критериев, основанных на коэффициентах корреляции можно лишь опровергнуть гипотезу независимости. При равенстве нулю коэффициента корреляции экспериментатор обладает лишь косвенной информацией в пользу независимости.

§ 4. При малых значениях n точные распределения ранговых коэффициентов связи можно найти перебрав все возможные расположения рангов и подсчитав частоты для каждого из значений рассматриваемого коэффициента. В целях сокращения вычислений можно наблюдения расположить в порядке возрастания по одному из признаков, тогда по каждому из остальных признаков можно получить $n!$ вариантов расположения рангов. Таким образом, придется перебрать „всего навсего“ $(n!)^{p-1}$ вариантов, что, конечно, возможно только с использованием вычислительной техники.

§ 4. Доказательства

Доказательство Свойств коэффициента Спирмена.

Так как коэффициент Спирмена вычисляется как обычный полный коэффициент корреляции, то он, например, равен 1 только в том случае, если обе переменные линейно связаны с положительным коэффициентом растяжения, т.е. с ростом одной переменной растет и другая. Поэтому если объекты упорядочены по возрастанию одного из признаков, то по второму признаку ранги также должны возрастать, т.е. совпадать с рангами первого признака.

⊗

Доказательство Теоремы II.1.

Очевидно, что средние значения обоих рядов совпадают:

$$x_{\cdot} = y_{\cdot} = \frac{1}{n} \sum_{i=1}^n i = \frac{n+1}{2}.$$

Аналогично совпадают и дисперсии:

$$d_x = d_y = \frac{1}{n} \sum_{i=1}^n i^2 - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$$

в силу известного соотношения

$$\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}. \quad (\text{II.3})$$

Сумма квадратов разностей рангов равна

$$\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 = d_x + d_y - 2 \frac{1}{n} \sum_{i=1}^n (x_i - x_{\cdot})(y_i - y_{\cdot}).$$

Подставляя полученные значения в формулу для корреляции

$$\text{Corr} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - x_{\cdot})(y_i - y_{\cdot})}{\sqrt{d_x d_y}},$$

получаем доказательство теоремы.

⊗

Доказательство Теоремы II.2.

Необходимо вычислить только среднее значение и дисперсию r_S — качественный факт асимптотической нормальности следует из центральной предельной теоремы.

Если признаки независимы, то

а) значения рангов (x_i, y_j) по разным признакам есть реализации независимых случайных величин;

б) значения рангов (x_i, x_j) по первому признаку (так же, как и по второму) есть реализации независимых одинаково распределенных случайных величин.

Так как ранги одинаково распределены, то

$$\mathbf{E} x_i = \frac{1}{n} \sum_{k=1}^n \mathbf{E} x_k = x. \quad \text{и} \quad \mathbf{E} y_i = \frac{1}{n} \sum_{k=1}^n \mathbf{E} y_k = y..$$

Следовательно, в силу независимости признаков, среднее

$$\mathbf{E} \frac{1}{n} \sum_{i=1}^n (x_i - x.)(y_i - y.) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(x_i - x.) \mathbf{E}(y_i - y.) = 0.$$

Отсюда и $\mathbf{E} r_S = 0$.

Для вычисления дисперсии r_S заметим, что

$$\sum_{i=1}^n (x_i - x.) \equiv 0$$

и все пары (x_i, x_j) одинаково распределены. Поэтому

$$\begin{aligned} 0 &= \mathbf{E} \left[(x_1 - x.) \sum_{i=1}^n (x_i - x.) \right] = \\ &= \mathbf{E}(x_1 - x.)^2 + \sum_{i=2}^n \mathbf{E}(x_1 - x.)(x_i - x.) = \\ &= \mathbf{E}(x_1 - x.)^2 + (n-1) \mathbf{E}(x_1 - x.)(x_2 - x.). \end{aligned}$$

Отсюда, $\mathbf{E}(x_1 - x.)(x_2 - x.) = -\mathbf{E}(x_1 - x.)^2 / (n-1)$. Заметив, что

$$\sum_{i=1}^n (x_i - x.)^2 \equiv n d_x,$$

получаем среднее значение $\mathbf{E}(x_1 - x.)^2 = d_x$.

Аналогичные формулы верны и для переменной y . Следовательно, положив $d_x = d_y = d$, получаем

$$\begin{aligned}
\mathbf{D} r_S &= \mathbf{E}(r_S)^2 = \frac{1}{d} \mathbf{E} \left(\sum_{i=1}^n (x_i - x.)(y_i - y.) \right)^2 = \\
&= \frac{1}{d} \sum_{i=1}^n \mathbf{E}(x_i - x.)^2 (y_i - y.)^2 + \\
&\quad + \frac{1}{d} \sum_{i,j \geq 2} \mathbf{E}(x_i - x.)(y_j - x.)(y_i - x.)(y_j - y.) = \\
&= \frac{1}{d} (n \mathbf{E}(x_1 - x.)^2 \mathbf{E}(y_1 - y.)^2 + \\
&\quad + n(n-1) \mathbf{E}(x_1 - x.)(x_2 - x.) \mathbf{E}(y_1 - y.)(y_2 - y.)).
\end{aligned}$$

Подставив полученные выше значения, получаем

$$\mathbf{D} r_S = 1/(n-1),$$

что и завершает доказательство теоремы.

⊗

Глава III. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Пытаясь побольше откусить от
свадебного каравая, подумай сначала,
сможешь ли ты всё это переварить?

Целью дискриминантного анализа является получение правил для классификации измерений в одну из нескольких известных совокупностей. Например, при поступлении больного в стационар первым делом производятся измерения большого числа характеристик организма пациента (СОЭ, артериальное давление, длительность болезни и пр.). По окончании лечения состояние пациента может быть охарактеризовано одной из следующих категорий: а) полностью нормализовалось, б) улучшилось, в) осталось без изменения, г) ухудшилось. Возникает естественный вопрос, нельзя ли попытаться предсказать результат лечения по значениям предварительных анализов и принятому методу лечения? Здесь прежде всего необходимо ответить на вопрос о принципиальной возможности построения такого правила. Для этого применяют методы многомерного дисперсионного анализа, которые позволяют оценить информативность как всего вектора рассматриваемых характеристик, так и каждой из его компонент и установить пригодность их для адекватного описания будущего состояния пациента.

§ 1. Дисперсионный анализ многомерных данных

Статистический эксперимент состоит в наблюдении над индивидами (объектами), разбитыми на J групп, причем с каждого индивида (объекта)

снимаются показания по p признакам. Пусть

$$x_{kji} - \text{результат измерения} \begin{cases} k\text{-ого признака } (k = \overline{1, p}) \text{ у} \\ i\text{-ого индивида } (i = \overline{1, n_j}) \text{ из} \\ j\text{-ой группы } (j = \overline{1, J}). \end{cases}$$

Общее число наблюдений $n = n_1 + \dots + n_J$.

Основное предположение.

Вектор всех p признаков $\vec{x}_{ji} = (x_{1ji}, \dots, x_{pji})'$, снятых с i -ого объекта в j -ой группе, есть рализация нормального случайного вектора

$$\vec{X}_j = (X_{1j}, \dots, X_{pj})' \sim \mathcal{N}_p(\vec{\mu}_j, \Lambda)$$

с вектором средних $\vec{\mu}_j$ и матрицей ковариаций $\Lambda > 0$ (одинаковой для всех групп).

Основная задача.

Проверить гипотезу о равенстве всех векторов средних значений

$$\mathbf{H} : \vec{\mu}_1 = \dots = \vec{\mu}_J. \quad (\text{III.1})$$

Точное решение эта задача имеет только в некоторых частных случаях, с рассмотрения которых мы и начнем. Введем следующие обозначения:

$$\vec{x}_{j\bullet} = \frac{1}{n_j} \sum_{i=1}^{n_j} \vec{x}_{ji} \quad - \quad \begin{array}{l} \text{вектор средних } j\text{-й группы} \\ \text{(несмещенная оценка } \vec{\mu}_j), \end{array}$$

$$\vec{x}_{\bullet\bullet} = \frac{1}{n} \sum_{j=1}^J n_j \vec{x}_{j\bullet} \quad - \quad \text{вектор общих средних,}$$

$$L_j = \sum_{i=1}^{n_j} \frac{(\vec{x}_{ji} - \vec{x}_{j\bullet})(\vec{x}_{ji} - \vec{x}_{j\bullet})'}{n_j} \quad - \quad \begin{array}{l} \text{ковариации } j\text{-й группы} \\ \text{(смещенная оценка } \Lambda \text{ по} \\ \text{данным } j\text{-й группы),} \end{array}$$

$$L = \frac{1}{n-J} \sum_{j=1}^J n_j L_j \quad - \quad \begin{array}{l} \text{общая матрица ковариаций} \\ \text{(несмещенная оценка } \Lambda \text{ по} \\ \text{всем данным).} \end{array}$$

1.1 Проверка гипотезы о центре распределения одной группы

Пусть в эксперименте наблюдается n реализаций p -мерного нормального вектора, о среднем значении которого $\vec{\mu}$ требуется проверить гипотезу

$$\mathbf{H} : \vec{\mu} = \vec{\mu}_0$$

с известным вектором $\vec{\mu}_0$. Например, при медицинских исследованиях задачу проверки эффективности лечения можно сформулировать как гипотезу о равенстве нулю ($\vec{\mu}_0 = \vec{0}$) вектора средних значений разностей всех контролируемых показателей до и после лечения.

Можно показать (аналогично тому, как это сделано при проверке гипотезы о коэффициенте корреляции), что статистика критерия отношения правдоподобия монотонно зависит от статистики

$$\mathbb{T}^2 = (n-1) (\vec{x} \cdot - \vec{\mu}_0)' L^{-1} (\vec{x} \cdot - \vec{\mu}_0),$$

где L – выборочная матрица ковариаций (смещенная). В монографии Т. Андерсона [1, стр.146-149] доказано, что соответствующим образом нормированная статистика $\mathbb{T}^2$ будет иметь распределение Фишера $\mathcal{F}_{p,n-p}$ с $(p, n-p)$ степенями свободы.

Теорема III.1. I) Если выборка получена из p -мерного нормального распределения со средним $\vec{\mu} = \vec{\mu}_0$, то статистика

$$\frac{n-p}{p(n-1)} \cdot \mathbb{T}^2 \sim \mathcal{F}_{p,n-p}.$$

II) Если $\hat{\mathbb{T}}^2$ – выборочное значение статистики $\mathbb{T}^2$, то критический уровень значимости критерия отношения правдоподобия при проверке гипотезы $\mathbf{H} : \vec{\mu} = \vec{\mu}_0$ о среднем значении нормального распределения равен

$$\alpha_{кр} = \mathbf{P}\{\mathbb{T}^2 > \hat{\mathbb{T}}^2\} = 1 - \mathcal{F}_{p,n-p} \left(\frac{n-p}{p(n-1)} \cdot \hat{\mathbb{T}}^2 \right). \quad (\text{III.2})$$

З 1. Доказательство этой теоремы целиком опирается на утверждения Теоремы I.13, стр.29. Другими словами, если k -мерный вектор $\vec{Z}$ имеет

нормальное распределение $\mathcal{N}_k(\vec{0}, B)$, а не зависящая от него случайная матрица $\mathbf{A}$ имеет распределение Уишарта $\mathbb{W}_k(B, \nu)$, то квадратичная форма

$$\frac{\nu - k + 1}{k} \vec{Z}' \mathbf{A}^{-1} \vec{Z} \sim \mathcal{F}_{k, \nu - k + 1} .$$

З 2. В одномерном случае ($p = 1$) статистика $\mathbb{T}^2$ совпадает с квадратом статистики Стьюдента, а критерий (III.2) с одновыборочным двусторонним критерием Стьюдента.

1.2 Проверка гипотезы об однородности двух совокупностей

Рассмотрим теперь задачу сравнения средних значений двух p -мерных нормальных совокупностей (групп). Легко видеть, что нормированная разность выборочных средних $\sqrt{n_1 n_2 / (n_1 + n_2)} (\vec{x}_{1\cdot} - \vec{x}_{2\cdot})$ имеет нормальное распределение $\mathcal{N}_p(\vec{0}, \Lambda)$. В соответствии с замечанием З1, для построения тестовой статистики нам необходимо найти оценку для матрицы Λ , имеющую распределение Уишарта. Поскольку распределение Уишарта обладает свойством аддитивности, то статистика

$$(n - 2)L = n_1 L_1 + n_2 L_2 \sim W_p(\Lambda, n - 2).$$

Таким образом,

$$\begin{aligned} \mathbb{T}^2 &= \frac{n_1 n_2 (n - p - 1)}{n (n - 2) p} (\vec{x}_{1\cdot} - \vec{x}_{2\cdot})' L^{-1} (\vec{x}_{1\cdot} - \vec{x}_{2\cdot}) \sim \\ &\sim \mathcal{F}_{p, n - p - 1}. \end{aligned} \quad (\text{III.3})$$

Теорема III.2. Если $\hat{\mathbb{T}}^2$ – выборочное значение статистики (III.3), то при проверке гипотезы $\mathbf{H} : \vec{\mu}_1 = \vec{\mu}_2$ о равенстве средних значений двух групп нормально распределенных данных критический уровень значимости

$$\alpha_{kp} = \mathbf{P}\{\mathbb{T}^2 > \hat{\mathbb{T}}^2\} = 1 - \mathcal{F}_{p, n - p - 1}(\hat{\mathbb{T}}^2) .$$

З 3. По этой же схеме можно осуществить построение критерия сравнения векторов средних значений J (> 2) групп, если проверяемая гипотеза связывает эти векторы посредством одной линейной функции – $\mathbf{H} : c_1 \vec{\mu}_1 + \dots + c_J \vec{\mu}_J = \vec{\mu}_0$. Пример такой задачи приведен в монографии [1, с.153], где сравниваются три сорта Iris'a ($J = 3$) по измерениям четырех характеристик лепестка и чашелистика ($p = 4$). Генетический анализ этих цветов дает некоторые основания предполагать, что векторы средних значений характеристик должны быть связаны соотношением $3\vec{\mu}_1 - 2\vec{\mu}_2 - \vec{\mu}_3 = \vec{0}$.

З 4. Приведенный здесь критерий предполагает, естественно, что степени свободы распределения Фишера положительны. Таким образом, общий объем выборки n должен быть не меньше числа признаков p плюс 2, где 2 есть не что иное, как число сравниваемых групп.

1.3 Проверка гипотезы однородности более двух групп

Рассмотрим теперь задачу проверки основной гипотезы (III.1, стр.69) для J (> 2) групп. Предположим, что число наблюдений $n = \sum n_j \geq p + J$ (см. предыдущее замечание).

Для построения критерия проверки гипотезы $\mathbf{H}$ воспользуемся результатами однофакторного дисперсионного анализа (см., например, [9], [11]) с одним измеряемым признаком ($p = 1$). Критерий отношения правдоподобия в этом случае основан на значениях статистики

$$\mathbb{F} = \frac{\frac{1}{J-1} \sum_{j=1}^J n_j (x_{j\cdot} - x_{\cdot\cdot})^2}{\frac{1}{n-J} \sum_{j=1}^J \sum_{i=1}^{n_j} (x_{ji} - x_{j\cdot})^2},$$

которая при справедливости гипотезы $\mathbf{H}$ имеет распределение Фишера с $g_1 = J - 1$ и $g_2 = n - J$ степенями свободы.

Рассмотрим многомерный аналог этой статистики

$$\mathbb{T}^2 = C \operatorname{tr}(QL^{-1}),$$

где

$$Q = \sum_{j=1}^J n_j (\vec{x}_{j\cdot} - \vec{x}_{\cdot\cdot}) (\vec{x}_{j\cdot} - \vec{x}_{\cdot\cdot})' - \quad (\text{III.4})$$

многомерный аналог числителя, L^{-1} - аналог знаменателя статистики $\mathbb{F}$. Статистика $\text{tr}(QL^{-1})$ называется статистикой Лоули-Хотеллинга.

Константу C нужно подобрать так, чтобы распределение $\mathbb{T}^2$ при гипотезе $\mathbf{H}$ имело более или менее привычный вид. Конечно, хотелось бы, чтобы статистика $\mathbb{T}^2$, как и в одномерном случае, имела распределение Фишера, однако, к сожалению, это недостижимо.

Известно, что при выборе $C = J - 1$ статистика $\mathbb{T}^2$ будет иметь асимптотическое ($n \rightarrow \infty$) хи-квадрат распределение $\mathcal{K}_\nu$ с $\nu = p(J - 1)$ степенями свободы. Это утверждение находится абсолютно в русле замечания $\mathbb{Z}23$, стр.37, поскольку статистика $\mathbb{T}^2$ есть не что иное, как логарифм статистики отношения правдоподобия, причем разность числа параметров при альтернативе и гипотезе как раз равна ν .

Мы воспользуемся здесь более точной „фишеровской“ аппроксимацией. А именно, выберем константу C из условия совпадения первых двух моментов статистики $\mathbb{T}^2$ с некоторым распределением Фишера. Таким образом, тестовая статистика (см. [2])

$$\mathbb{T}^2 = C \sum_{j=1}^J n_j (\vec{x}_{j\cdot} - \vec{x}_{\cdot\cdot})' L^{-1} (\vec{x}_{j\cdot} - \vec{x}_{\cdot\cdot}) \quad (\text{III.5})$$

будет иметь приближенное распределение Фишера $\mathcal{F}_{g_1, g_2}$ с (g_1, g_2) степенями свободы, где константа

$$C = \frac{n - J - p - 1}{(n - J)(J - 1)p} \frac{g_2}{g_2 - 2},$$

а параметры g_1 и g_2 определяются следующим образом:

$$\left. \begin{aligned} g_1 &= p(J - 1) \frac{n - J - p}{n - J - p - (J - 2)(p - 1)} \\ g_2 &= n - J - p + 1 \end{aligned} \right\}, \quad \text{если} \quad n - (J - 1)p - 2 > 0,$$

если же $n - (J - 1)p - 2 \leq 0$, то

$$g_1 = +\infty, \quad g_2 = (n - J - p + 1) - \frac{(n - J - p - 1)(n - J - p - 3)(n - (J - 1)p - 2)}{(n - J - 1)(n - p - 2)}.$$

Теорема III.3. Если $\hat{\mathbb{T}}^2$ – выборочное значение статистики (III.5), то при проверке гипотезы $\mathbf{H} : \vec{\mu}_1 = \dots = \vec{\mu}_J$ о равенстве средних значений J нормальных совокупностей с одинаковыми во всех группах ковариациями

$$\alpha_{кр} = \mathbf{P}\{\mathbb{T}^2 > \hat{\mathbb{T}}^2\} \approx 1 - \mathcal{F}_{g_1, g_2}(\hat{\mathbb{T}}^2). \quad (\text{III.6})$$

З 5. Как при различении двух групп ($J = 2$), так и в одномерном случае ($p = 1$) предложенный здесь критерий полностью идентичен построенным в предыдущих разделах критериям Фишера. В обоих этих случаях равенство (III.6) будет точным.

З 6. Когда число степеней свободы g_1 полагается равным $+\infty$, распределение Фишера определяется обратной хи-квадрат случайной величиной с g_2 степенями свободы. Легко вывести, что в этом случае

$$\mathcal{F}_{\infty, g_2}(t) = 1 - \mathcal{K}_{g_2}\left(\frac{g_2}{t}\right),$$

и α -квантиль распределения Фишера $\mathcal{F}_{\infty, g_2}^{(\alpha)} = g_2 / \mathcal{K}_{g_2}^{(1-\alpha)}$.

З 7. На практике применяют также и другие критерии сравнения групп. В большинстве своем эти критерии основаны на $t = \min(J - 1, p)$ ненулевых собственных числах $\lambda_1, \dots, \lambda_t$ матрицы QL^{-1} . Наиболее популярен из них критерий Λ -Уилкса. Тестовая статистика этого критерия равна

$$\Lambda = \prod_{k=1}^t \frac{1}{1 + \lambda_k}.$$

Точное распределение статистики Λ известно только при $t \leq 2$. При других значениях t применяют асимптотическое распределение:

$$-\left(n - \frac{p + J + 2}{2}\right) \ln \Lambda \rightsquigarrow \mathcal{K}_{p(J-1)}, \quad n \rightarrow \infty.$$

Парные и множественные сравнения групп

Критерий (III.6) позволяет сделать вывод о различиях, существующих между J группами в целом, но он не дает информации о том, насколько определенные группы отличаются одна от другой. Обычная практика состоит в том, что после отвержения основной гипотезы проводят попарное сравнение всех групп данных с помощью того же самого критерия. Однако при этом возникают абсолютно такие же проблемы, как и при построении доверительных утверждений сразу для нескольких параметров: многократное применение критерия оставляет открытым вопрос доверия к полученным результатам (см. замечание Z15, стр.20). К настоящему времени разработано несколько способов решения этой задачи. Мы опишем здесь так называемый S-метод, предложенный Шеффе. Следуя этому методу, необходимо, во-первых, оценку ковариационной матрицы L производить по всем выборочным данным. Во-вторых, для проверки гипотезы $\mathbf{H}_{u/v} : \vec{\mu}_u = \vec{\mu}_v$ о равенстве средних значений признаков в группах u и v в качестве тестовой статистики взять статистику (с теми же самыми g_1, g_2)

$$\mathbb{T}_{u/v}^2 = \frac{n - J - p - 1}{p(n - J)} \frac{g_2}{(g_2 - 2)} \frac{n_u n_v}{(n_u + n_v)} (\vec{x}_{u\cdot} - \vec{x}_{v\cdot})' L^{-1} (\vec{x}_{u\cdot} - \vec{x}_{v\cdot}).$$

В-третьих, критический уровень значимости вычислять через распределение Фишера по формуле

$$\alpha_{\text{кр}} = 1 - \mathcal{F}_{g_1, g_2} \left(\frac{1}{J-1} \hat{\mathbb{T}}_{u/v}^2 \right). \quad (\text{III.7})$$

Z 8. Если здесь опустить делитель $J - 1$, то получим критический уровень значимости для проверки гипотезы $\mathbf{H}_{u/v}$ для фиксированной пары совокупностей (u, v) . Применяя (III.7), мы уменьшаем вероятность отвержения гипотезы $\mathbf{H}_{u/v}$ (тем самым ужесточая критерий), что позволяет достигнуть номинального уровня значимости при множественном сравнении всех возможных пар.

§ 2. Классифицирующие правила

После того как методами дисперсионного анализа было установлено различие между группами данных, возникает задача построения правила, относящего вновь полученные данные к той или иной группе. Кроме того, необходимо выработать меру, оценивающую информационное содержание одного или нескольких измеряемых признаков.

Дистант совокупности признаков

Меру разделительной способности признаков естественно связать с тестовой статистикой (III.5).

О п р е д е л е н и е 1. Статистика

$$T^2(\vec{X}) = \frac{1}{n - J} \sum_{j=1}^J n_j (\vec{x}_{j\cdot} - \vec{x}_{\cdot\cdot})' L^{-1} (\vec{x}_{j\cdot} - \vec{x}_{\cdot\cdot}) \quad (\text{III.8})$$

называется многомерным дистантом p признаков $\vec{X} = (X_1, \dots, X_p)$.

§ 9. В рамках проблемы классификации (кластеризации) статистику $(n - J) T^2$ принято называть расстоянием Махалобиса.

Рассмотрим некоторые свойства многомерного дистанта.

(I) Дистант инвариантен относительно невырожденных линейных преобразований признаков. Иными словами, имеет место тождество

$$T^2(Z_1, \dots, Z_p) = T^2(X_1, \dots, X_p)$$

при всех $\vec{Z} = U' \vec{X}$ с невырожденной матрицей U . Это легко устанавливаемое свойство весьма важно, поскольку придает некоторую уверенность в независимости от произвола в выборе признаков.

(II) Дистант монотонно зависит от пространства признаков:

$$T^2(X_1) \leq T^2(X_1, X_2) \leq \dots \leq T^2(X_1, \dots, X_p).$$

Дискриминантные признаки

Существуют преобразования, понижающие размерность пространства признаков, но не меняющие многомерный дистант.

О п р е д е л е н и е 2. Положим

$$\vec{d}_j = S^{-1}(\vec{x}_{j\cdot} - \vec{x}_{\cdot\cdot}), \quad j = \overline{1, J} \quad (\text{III.9})$$

и определим $(p \cdot J)$ -матрицу $D = (\vec{d}_1, \dots, \vec{d}_J)$. Переменные

$$(V_1, \dots, V_J) = (\vec{d}'_1 \vec{X}, \dots, \vec{d}'_J \vec{X}) = (D' \vec{X})'$$

называют элементарными дискриминантными признаками.

Т е о р е м а III.4. *Значение дистанта не изменяется при переходе к элементарным дискриминантным признакам*

$$T^2(V_1, \dots, V_J) = T^2(X_1, \dots, X_p).$$

§ 10. Элементарные дискриминантные признаки не являются линейно-независимыми, так как

$$\sum_{j=1}^J n_j V_j = 0.$$

Таким образом, если число групп $J \leq p$, то число признаков, подлежащих рассмотрению, сокращается, по крайней мере, до $J - 1$.

Каждому из элементарных дискриминантных признаков можно придать содержательную интерпретацию. А именно, признак V_j можно использовать для выделения группы j из множества всех остальных групп.

Как отмечалось в свойстве (I), дистант не изменяется при невырожденных линейных преобразованиях. Представляет интерес отыскание линейной комбинации признаков, обладающей наилучшей разделяющей способностью не только по всем p признакам, но и при переходе к меньшему числу признаков.

Неэлементарные дискриминантные признаки

О п р е д е л е н и е 3. Вектор

$$\vec{W} = (W_1, \dots, W_t)' = E' \vec{X}$$

с нормированными столбцами матрицы преобразования E называется вектором неэлементарных дискриминантных признаков, если

- (i) W_1 имеет максимальное значение дистанта $T^2(W_1)$ среди всех нормированных линейных комбинаций $\vec{X}$;
- (ii) (W_1, W_2) имеет максимальное значение дистанта $T^2(W_1, W_2)$ среди всех некоррелированных пар нормированных линейных комбинаций $\vec{X}$;
- (iii) и т.д. ...

Свойства неэлементарных дискриминантных признаков.

1) Столбцы матрицы $E = (\vec{e}_1, \dots, \vec{e}_t)$ совпадают с нормированными собственными векторами обобщенной проблемы собственных значений

$$\frac{1}{n - J} H \vec{e}_k = \lambda_k S \vec{e}_k \quad (k = 1, \dots, t),$$

причем все собственные числа отличны от нуля и упорядочены по убыванию $(\lambda_1 \geq \dots \geq \lambda_t)$.

2) Число неэлементарных дискриминантных признаков равно

$$t = \text{rang}(H) = \min(p, J - 1).$$

3) Ковариационная матрица $\text{Cov}(\vec{W}) = \mathbf{I}$, то есть дисперсии неэлементарных признаков единичны и внутри отдельных групп не коррелированы.

4) Дистант любой совокупности неэлементарных дискриминантных признаков равен сумме соответствующих характеристических чисел

$$T^2(W_{i_1}, \dots, W_{i_m}) = \lambda_{i_1} + \dots + \lambda_{i_m}.$$

5) Неэлементарные дискриминантные признаки порождают то же линейное пространство, что и элементарные дискриминантные признаки,

то есть максимальная линейно-независимая часть элементарных признаков связана с неэлементарными признаками посредством невырожденного линейного преобразования.

Классификация

Теперь мы уже можем приступить к основной части дискриминантного анализа – разделению вновь полученных данных на группы. Метод классификации, рассматриваемый здесь, основан на критерии значимости.

Пусть для некоторого индивида получен вектор данных $\vec{x} = (x_1, \dots, x_p)'$ и вычислен вектор признаков $\vec{w} = (w_1, \dots, w_t)'$. Как мы выяснили выше, степень различия между двумя группами может быть охарактеризована величиной статистики $T_{u/v}$ (см. стр.75). Если формально считать значение $\vec{w}$ принадлежащим некой „нулевой“ группе, то за расстояние от этой группы до j -ой группы можно взять величину

$$k_j = T_{0/j} = \frac{n - J - t + 1}{t(n - J)} \frac{n_j}{n_j + 1} (\vec{w} - \vec{w}_{j\cdot})'(\vec{w} - \vec{w}_{j\cdot}),$$

где $\vec{w}_{j\cdot}$ – вектор средних значений дискриминантных признаков в j -ой группе. Если новые данные получены из j -ой группы, то распределение k_j приблизительно совпадает с распределением Фишера с $(t, n - J - t + 1)$ степенями свободы.

Таким образом, если

$$k_j \leq \mathcal{F}_{t, n-J-t+1}^{(\alpha)}, \quad (\text{III.10})$$

то вновь полученные данные можно отнести к группе j .

§ 11. При построении меры близости данных к j -ой группе было учтено, что размерность пространства дискриминантных признаков $\vec{w}$ равна t (p заменено на t) и матрица ковариаций $S_W = \mathbf{Cov}(\vec{W}) = \mathbf{I}$ – единичная.

Может так случиться, что неравенство (III.10) будет выполняться сразу для нескольких групп. Чтобы избежать такой многозначной классификации, этот способ часто модернизируют и относят нового индивида к

группе с наименьшим значением k_j . Кроме того, указанную классификацию можно уточнить, если дополнительно учесть априорные вероятности P_j попадания индивидов в группы, которые либо известны из неких дополнительных исследований, либо оцениваются по выборке. В этом случае применяется коэффициент

$$l_j = \left(1 + \frac{t}{n - J - t + 1} k_j \right)^{\frac{n+1}{2}} \frac{1}{P_j}.$$

Индивид при этом относится к группе с наименьшим l_j .

§ 12. Предложенное уточнение метода классификации получается, если относить индивида к группе, для которой совокупность выборочных данных и дополнительный вектор наблюдений индивида дают наибольшее значение функции правдоподобия.

Классификация не по всем дискриминантным признакам

Свойство 4) неэлементарных дискриминантных признаков дает основание для подразделения совокупности дискриминантных признаков по степени разделяющей способности. Более того, представление

$$T^2(W_1, \dots, W_t) = \lambda_1 + \dots + \lambda_t = T^2(W_1, \dots, W_s) + \lambda_{s+1} + \dots + \lambda_t$$

позволяет отбросить малоинформативные признаки и не использовать их при классификации. Особенный интерес представляет сокращение числа признаков до одного или двух, поскольку в этих случаях неравенства (III.10) легко интерпретируются геометрически (интервалы на числовой прямой или круги на плоскости).

Отбрасывание излишних дискриминантных признаков можно производить по критерию значимости. Для этого используем статистику

$$\mathbb{T} = \frac{n - J - p + 1}{s(p + J - s - 1)} (\lambda_1 + \dots + \lambda_s),$$

которая имеет приближенно распределение Фишера. Однако чаще всего для оценки разделяющей способности k последних неэлементарных признаков

используют статистику Лямбда-Уилкса

$$\Lambda_k = \prod_{i=t-k+1}^t \frac{1}{1 + \lambda_i} = \frac{1}{1 + \lambda_t} * \dots * \frac{1}{1 + \lambda_{t-k+1}}, \quad k = t, \dots, 1.$$

При больших n величина

$$X_k^2 = - \left(n - 1 - \frac{p+J}{2} \right) \ln \Lambda_k$$

приближенно распределена по закону хи-квадрат с $(p - t + k)(J - t - 1 + k)$ степенями свободы. В частности, разделяющая способность всех t дискриминантных признаков будет значимой, если

$$1 - \mathcal{K}_{p(J-1)} \left(- \left(n - 1 - \frac{p+J}{2} \right) \ln \Lambda_t \right) \leq \alpha,$$

где $\mathcal{K}_\nu$ – функция распределения хи-квадрат с ν степенями свободы (сравните с Z6, стр.74).

Другой способ оценки реальной полезности дискриминантной функции можно получить, рассматривая коэффициент канонической корреляции

$$R_j^* = \sqrt{\frac{\lambda_j}{1 + \lambda_j}},$$

который является мерой связи между j -ой дискриминантной функцией и группами. Чем ближе этот коэффициент к 1, тем выше зависимость групп от значений дискриминантной функции, и наоборот, чем меньше R_j^* , тем хуже распознаются группы по значениям W_j .

Оценка дискриминантной ошибки

Качество правил классификации можно оценить несколькими способами.

Простейший состоит в подстановке в эти правила выборочных данных. Однако понятно, что такой способ весьма приближенно отражает истинную картину и дает чаще всего чересчур оптимистичный результат: доли ошибок получаются незначительными.

Другой способ получения несмещенной оценки ошибки состоит в разделении выборки на две части, по одной из которых строят правило классификации, а другую используют как контрольную для оценки ошибок классификации. Этот метод может быть применен лишь в редких случаях, когда

объем исходной выборки велик и уменьшение его вдвое не слишком сильно отражается на результатах.

Метод, идейно близкий к последнему, предложил П.А.Лахенбрух. Этот метод состоит из n шагов (по общему числу наблюдений), на каждом из которых в контрольную группу включается только по одному наблюдению. Доля неправильно классифицированных векторов наблюдений дает оценку вероятности ошибок решающего правила. При больших объемах выборок этот способ приводит к результатам, почти совпадающим с первым простейшим способом.

Исключение излишних признаков

Для применения дискриминантного анализа особое значение приобретают высокоэффективные сочетания наблюдаемых признаков. Для построения правила сокращения признаков рассмотрим так называемые показатели необходимости

$$U_i = T^2(X_1, \dots, X_p) - T^2(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p).$$

Можно показать, что

$$U_i = \frac{1}{(n-J)t_{ii}} \sum_{j=1}^J n_j d_{ij}^2,$$

где d_{ij} – весовые коэффициенты элементарных дискриминантных признаков (III.9, стр.77), а t_{ii} – диагональные элементы матрицы S^{-1} . Пошаговая процедура исключения признаков признает избыточным признак с наименьшим показателем необходимости. Окончание процесса исключения можно поставить в зависимость от того, насколько отдельные признаки дают значимое увеличение дистанта. Таким образом, процесс сокращения можно продолжать до тех пор, пока все остающиеся признаки $(X_{j_1}, \dots, X_{j_{p^*}})$ не будут удовлетворять неравенству

$$\frac{n - J - p^* + 1}{J - 1} \frac{U_i}{1 + T^2(X_{j_1}, \dots, X_{j_{p^*}}) - U_i} > \mathcal{F}_{J-1, n-J-p^*+1}^{(\alpha)}.$$

Глава IV.

ПРИМЕРЫ ПРИМЕНЕНИЯ МЕТОДОВ МНОГОМЕРНОГО АНАЛИЗА

Во всех отношениях приятнее учиться
на ошибках других людей, чем на их
достижениях.

Пример I. *Исследование связи между ростом и весом женщин.*

Изучалась связь между ростом (в дюймах – X) и весом (в фунтах – Y) у 4995 женщин Великобритании. Конечно, если быть до конца последовательным, то следовало бы данные записать в виде 4995 строк. Однако по неизвестным нам причинам данные были сгруппированы, то есть истинные измеренные значения характеристик были заменены центрами интервалов группировки. Для данных такого сорта можно сохранять только частоты индивидов, попавших в ту или иную двумерную ячейку. Из приведенной ниже таблицы видно, что 1454 женщины имели рост от 63 до 65 дюймов, при этом у 207 из них вес колебался в пределах от 125.5 до 131.5 фунтов.

Первые выборочные моменты равны

	Среднее	Дисперсия	Стандартное отклонение
Рост	$m_x = 63.06$	$s_x^2 = 7.249$	$s_x = 2.692$
Вес	$m_y = 132.82$	$s_y^2 = 507.457$	$s_y = 22.527$

Оценки центральных моментов $a_{km} = \mathbf{E}[(X - \mu_x)^k(Y - \mu_y)^m]$ (только те, которые нам понадобятся при статистическом анализе) равны

$m \backslash k$	0	1	2	3	4
0			7.249		166.901
1		19.523		499.294	
2	507.457		4685.44		
3		26754.2			
4	1250330				

Распределение веса и роста женщин Великобритании, 1951г.

Y	X											Σ
	54	56	58	60	62	64	66	68	70	72	74	
278.5						1						1
272.5												0
266.5						1						1
260.5							1					1
254.5												0
248.5					1	1						2
242.5							1					1
236.5							1					1
230.5					2				1			3
224.5					1	2	1					4
218.5			1		2	1		1				5
212.5				2	1	6		1	1			11
206.5				2	2	3	2		1			10
200.5			4	2	6	2						14
194.5				1	3	7	7	4	1			23
188.5			1	5	14	8	12	3	1	2		46
182.5			1	7	12	26	9	5		1	2	63
176.5			5	8	18	21	15	11	7		2	87
170.5			2	11	17	44	21	13	3	1		112
164.5		1	3	12	35	48	30	15	5	3		152
158.5			8	17	52	42	36	21	9			185
152.5		1	7	30	81	71	58	21	2	2		273
146.5		2	13	36	76	91	82	36	8	1		345
140.5		1	6	55	101	138	89	50	8			448
134.5			15	64	95	175	122	45	5			521
128.5		1	19	73	155	207	101	25	3			584
122.5		3	34	91	168	200	81	12	1	1		591
116.5		3	24	108	184	184	50	8				561
110.5		5	33	119	165	124	22	4				472
104.5	1	3	33	87	95	35	6					260
98.5	2	5	29	59	45	16	3					159
92.5		6	10	21	9							46
86.5		1	5	3								9
80.5	2	1	1									4
Σ	5	33	254	813	1340	1454	750	275	56	11	4	4995

Коэффициент корреляции Пирсона.

Оценка полного коэффициента корреляции

$$r = \frac{19.523}{2.692 \cdot 22.527} = 0.322.$$

Соответствующее ему корреляционное отношение равно

$$r^2 \cdot 100\% = 10.4\%.$$

Другими словами, только 10.4% изменчивости веса женщин может быть объяснено влиянием на него показателя роста.

Оценка асимптотической дисперсии r (I.2, стр.10) (без делителя $\sqrt{n}$)

$$\hat{\tau}_\rho^2 = 0.902.$$

Воспользовавшись нормальной аппроксимацией, отсюда легко построить приближенный 95%-й доверительный интервал для ρ :

$$\left(0.322 \mp 1.960 \sqrt{\frac{0.902}{4995}} \right) = (0.296; 0.348),$$

где число 1.960 есть ни что иное, как верхняя 2.5%-квантиль стандартного нормального закона.

Если есть некоторые основания считать наши данные реализациями нормального случайного вектора, то доверительный интервал для ρ можно построить воспользовавшись преобразованием Фишера (см. описание на стр.17):

$$z = \frac{1}{2} \ln \left(\frac{1 + 0.322}{1 - 0.322} \right) = 0.334,$$

$$\tanh \left(0.334 \mp \frac{1.960}{\sqrt{4992}} \right) = (0.297; 0.347).$$

Доверительный интервал для ρ не содержат точку 0. Отсюда можно сделать вывод, что на 5%-ом уровне значимости гипотеза о равенстве нулю коэффициента корреляции (при двухсторонней альтернативе) должна быть отвергнута. Заметим однако, что при вычислении асимптотической дисперсии $\hat{\tau}_\rho^2$ по формуле (I.2, стр.10) вместо неизвестного коэффициента корреляции нами была подставлена его оценка $\hat{\rho} = r$. Если же проверяется гипотеза $\mathbf{H} : \rho = 0$, то дисперсию $\hat{\tau}_\rho^2$ следует вычислять при значении $\rho = 0$:

$$\hat{\tau}_\rho^2 = \frac{a_{22}}{s_x^2 s_y^2} = \frac{4685.44}{7.249 \cdot 507.457} = 1.274.$$

Более того, если требуется проверить гипотезу независимости значений роста и веса, то как отмечено в замечании 29, стр.13, следует положить

$$\hat{\tau}_\rho^2 = 1.$$

Вы будете смеяться, но при таком огромном объеме выборки столь слабое расхождение в значении дисперсии не играет никакой существенной роли. Во всех трех случаях критический уровень значимости будет гораздо ниже 10^{-50} :

$$\alpha_{\text{кр}} = 2 - 2\Phi\left(\sqrt{\frac{n}{\hat{\tau}_\rho^2}} |r|\right) < 2(1 - \Phi(20)) < 0.5 \cdot 10^{-90}.$$

Среднеквадратическая регрессия.

Теперь перейдем к задаче построения наилучшего прогноза одной из переменных (скажем, веса Y) по другой переменной. Сначала построим линейную регрессию (ЛСКР) Y на X .

Коэффициент линейной регрессии Y на X равен

$$\hat{\beta} = \frac{a_{11}}{s_x^2} = \frac{19.523}{7.249} = 2.693.$$

Его асимптотическая дисперсия (в предположении нормальности) вычисляется по формуле (I.7, стр.18)

$$\hat{\tau}_\beta^2 = \frac{507.457}{7.249}(1 - 0.322^2) = 47.47, \quad \sqrt{\frac{\hat{\tau}_\beta^2}{n}} = 0.0975.$$

Асимптотический 95%-доверительный для β интервал равен

$$(2.693 \mp 1.960 \cdot 0.0975) = (2.502; 2.884).$$

Для сравнения вычислим асимптотическую дисперсию $\hat{\tau}_\beta^2$ по формуле (I.6, стр.18) без предположения нормальности наблюдений:

$$\hat{\tau}_\beta^2 = \frac{4685.44 - 2 \cdot 499.294 \cdot 2.693 + 166.901 \cdot 2.693^2}{7.249^2} = 61.02, \quad \sqrt{\frac{\hat{\tau}_\beta^2}{n}} = 0.1105.$$

Различие не существенное, благодаря гигантскому объему выборки.

Уравнение ЛСКР Y на X (наилучший линейный прогноз веса женщины по её росту) имеет вид

$$y_*(x) = 132.82 + 2.693(x - 63.06) = 2.693x - 37.003.$$

Остаточная дисперсия ЛСКР равна

$$s_{y_*}^2 = 507.457(1 - 0.322^2) = 454.841.$$

Асимптотическая дисперсия ЛСКР в точке x (I.8, стр.19) (снова без делителя $\sqrt{n}$) равна

$$\hat{\tau}_y^2(x) = s_y^2(1 - r^2) \left(1 + \left(\frac{x - m_x}{s_x} \right)^2 \right) = 454.88 + 62.75(x - 63.06)^2.$$

Таким образом, 95%-доверительная полоса Уоркинга-Хотеллинга

$$\left(y_*(x) \pm C_n^{(\alpha)} \sqrt{\frac{\hat{\tau}_y^2(x)}{n}} \right) = \left(2.693x - 37.003 \mp 0.035 \sqrt{454.88 + 62.75(x - 63.06)^2} \right),$$

где константа $C_n^{(\alpha)} = \sqrt{2\mathcal{F}_{2,4993}^{(0.05)}} = 2.448$.

Поскольку наши данные изначально группированы, поэтому для них вполне естественно попытаться построить (нелинейную) среднеквадратическую регрессию Y на X . Эта регрессия равна условному среднему Y при фиксированном X :

j	x_j	$m_y(x_j)$	n_j
1	54	92.5	5
2	56	111.41	33
3	58	122.05	254
4	60	124.43	813

j	x_j	$m_y(x_j)$	n_j
5	62	130.22	1340
6	64	134.59	1454
7	66	140.48	750
8	68	146.37	275

j	x_j	$m_y(x_j)$	n_j
9	70	157.32	56
10	72	163.41	11
11	74	179.50	4

Дисперсия условных средних (при числе групп $k = 11$) равна

$$s_{y*}^2 = \sum_{j=1}^k \frac{n_j}{n} (m_y(x_j) - m_y)^2 \equiv \sum_{j=1}^k \frac{n_j}{n} m_y(x_j)^2 - m_y^2 = 54.199.$$

Корреляционное отношение, вычисляемое как отношение дисперсии условных средних к безусловной дисперсии Y -ов (см. [10, стр.26, теорема I.10]), равно

$$e_y^2 = \frac{s_{y*}}{s_y^2} = \frac{54.199}{507.457} = 0.107.$$

Разность между корреляционным отношением и квадратом коэффициента корреляции служит мерой „нелинейности“ регрессии. Известно, что если регрессия линейна и внутри групп наблюдения имеют нормальное распределение, то статистика

$$\frac{n - k}{k - 2} \frac{e_y^2 - r^2}{1 - e_y^2} \sim \mathcal{F}_{k-2, n-k}.$$

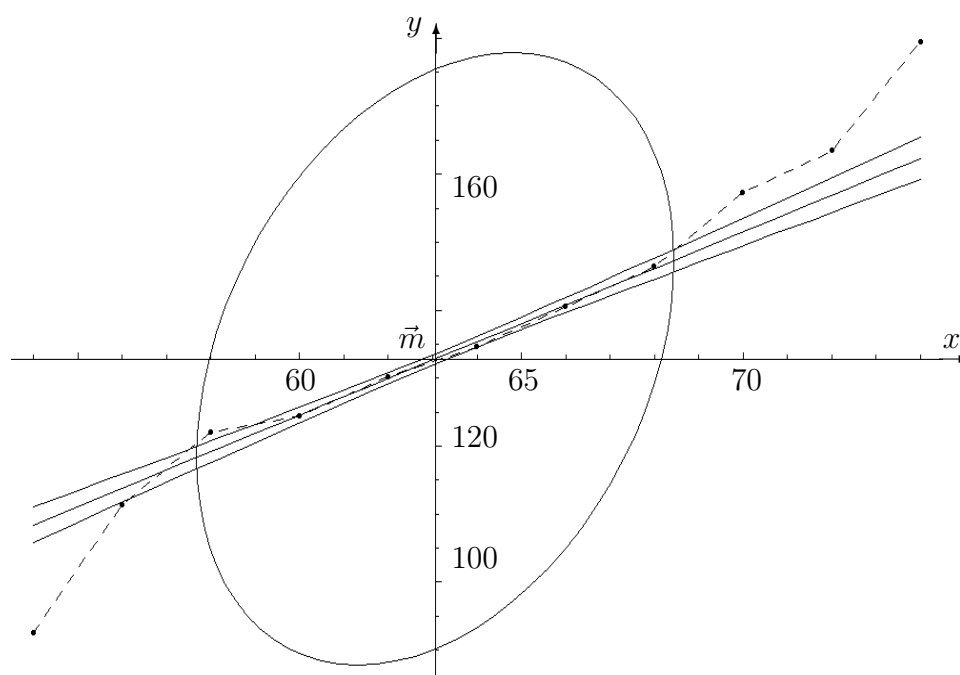
Для наших данных значение указанной статистики равно 1.98, поэтому кри-

тический уровень значимости при проверке гипотезы линейности регрессии

$$\alpha_{\text{кр}} = 1 - \mathcal{F}_{9,4884}(1.98) = 0.04 \quad -$$

гипотеза линейности должна быть принята. Этот результат подтверждает визуальное сравнение графиков обеих регрессий, приведенных на следующем рисунке (пунктирная линия – график СКР). Кроме того, этот рисунок содержит линии доверительной полосы Уоркинга-Хотеллинга, а также эллипс рассеяния (см. [10, (I.6), стр.23]). Полоса Уоркинга-Хотеллинга получилась очень узкой снова благодаря большому объёму выборки.

Широкая область, охватываемая эллипсом, говорит о плохих возможностях прогноза значения веса женщин по значениям их роста. Аналогичный результат был получен нами при рассмотрении коэффициентов детерминации.



Выводы.

- 1) Связь между ростом и весом женщин статистически высоко значима.
- 2) Среднеквадратическая регрессия линейна.
- 3) Прогностические качества регрессии близки к нулю. Высокая статистическая значимость объясняется исключительно большим объемом наблюдений.

Пример II. Исследование влияния на урожайность погодных условий.

В одной из областей Англии в течение 20 лет собирались данные об урожайности сена (в центнерах – X_1), весеннем количестве осадков (в дюймах – X_2) и накопленной за весну температурой выше 42°F (X_3). В иллюстративных целях будем считать, что вектор (X_1, X_2, X_3) имеет нормальное распределение.

Результаты первичного статистического анализа собраны в следующих двух таблицах.

Признак	Среднее	Дисперсия	Стандартное отклонение
Урожай	$m_1 = 28.02$	$s_1^2 = 19.54$	$s_1 = 4.42$
Осадки	$m_2 = 4.91$	$s_2^2 = 1.21$	$s_2 = 1.10$
t	$m_3 = 594$	$s_3^2 = 7225$	$s_3 = 85$

Матрица корреляций

$$\mathcal{R} = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline X_1 & 1 & 0.80 & -0.40 \\ X_2 & 0.80 & 1 & -0.56 \\ X_3 & -0.40 & -0.56 & 1 \end{array}$$

Гипотезы независимости.

Сначала проверим гипотезу о независимости всех трех признаков, воспользовавшись утверждением следствия 3, стр.38. Статистика критерия

$$\mathfrak{T} = -\ln(|\mathcal{R}|) = 1.41.$$

Следовательно, критический уровень значимости (по распределению хи-квадрат с $\nu = \frac{3 \cdot 2}{2} = 3$ степенями свободы и $\eta = 20 - \frac{2 \cdot 3 + 11}{6} = 17.2$)

$$\alpha_{\text{кр}} = 1 - \mathcal{K}_3(17.2 \cdot 1.41) = 0.00002 \quad -$$

гипотеза независимости должна быть отвергнута.

Корреляция между урожайностью и температурой $r = -0.4$ слабо значима (по критерию Стьюдента для корреляции с $20-2=18$ ст. св.):

$$t = \sqrt{18} \frac{0.4}{\sqrt{1 - 0.4^2}} = 1.851, \quad \alpha_{\text{кр}} = 2(1 - \mathcal{S}_{18}(1.851)) = 0.08.$$

Не смотря на этот факт, всё равно несколько странно видеть, что с ростом температуры средняя урожайность сена уменьшается. Попытаемся исправить ситуацию, подсчитав частную корреляцию (см. [10, стр. 13, (I.5)]):

$$r_{13*2} = \frac{-0.40 - (0.80)(-0.56)}{\sqrt{(1 - 0.80^2)(1 - 0.56^2)}} = 0.097.$$

Частная корреляция между урожайностью и температурой, во-первых, стала положительной (Уф! ($\widehat{\odot\odot}$)), во-вторых, незначимой (по критерию Стьюдента для частной корреляции с 17 ст. св.):

$$t = \sqrt{17} \frac{0.097}{\sqrt{1 - 0.097^2}} = 0.40, \quad \alpha_{\text{кр}} = 2(1 - \mathcal{S}_{17}(0.40)) = 0.69.$$

Линейная среднеквадратическая регрессия.

Для построения регрессии X_1 на (X_2, X_3) найдем матрицу ковариаций и обратную к ней

$$L = S \mathcal{R} S = \begin{pmatrix} 19.54 & 3.89 & -150.28 \\ 3.89 & 1.21 & -52.36 \\ -150.28 & -52.36 & 7225 \end{pmatrix},$$

где S – диагональная матрица стандартных отклонений,

$$L^{-1} = \begin{pmatrix} 0.144 & -0.484 & -0.00052 \\ -0.484 & 2.836 & 0.10 \\ -0.00052 & 0.10 & 0.0002 \end{pmatrix}.$$

Таким образом, коэффициенты линейной регрессии (см. формулу в последнем абзаце на стр.7 [10])

$$b_{12*3} = -\frac{-0.484}{0.144} = 3.36; \quad b_{13*2} = -\frac{-0.00052}{0.144} = 0.00364.$$

Уравнение регрессии имеет вид

$$x_{1*} = 28.02 + 3.37(x_2 - 4.91) + 0.00364(x_3 - 594) = 3.37x_2 + 0.00364x_3 + 9.311.$$

Качество прогноза этой функции характеризуют остаточная дисперсия и множественный коэффициент корреляции:

$$s_{1*}^2 = \frac{1}{0.144} = 6.944, \quad r_{1*} = \sqrt{1 - \frac{6.944}{19.54}} = 0.803.$$

Проверку гипотезы об отсутствии линейной связи между урожайностью и обоими погодными факторами (гипотезу о равенстве нулю множественного коэффициента корреляции) проверим по критерию, описанному в теореме I.21, стр.39:

$$t = \frac{20 - 3}{3 - 1} \frac{0.80^2}{1 - 0.80^2} = 15.11, \quad \alpha_{\text{кр}} = 1 - \mathcal{F}_{2,17}(15.11) = 0.0002.$$

Таким образом, наблюдается высокая статистическая значимость влияния погодных условий на урожайность сена. Степень этого влияния характеризуется корреляционным отношением

$$r_{1*}^2 \cdot 100\% = 64.5\%.$$

Если вспомнить, что квадрат полного коэффициента корреляции $r_{12}^2 = 0.64$, то отсюда можно сделать вывод, что это влияние почти полностью обусловлено весенней температурой.

Выводы.

I) Влияние погодных условий на урожайность сена статистически высоко значимо.

II) Весенняя температура оказывает влияние на урожайность исключительно посредством влияния её на количество осадков. С повышением температуры скорее всего следует ожидать уменьшения осадков, что в свою очередь может повлечь понижение урожайности сена.

III) Урожайность сена на 64% объясняется весенней температурой и только на 0.5% количеством весенних осадков.

Пример III. Экономические показатели в промышленности.

На 50 промышленных предприятиях, производящих радиоэлектронную продукцию, изучалось распределение следующих пяти признаков:

X_1 – средняя заработная плата;

X_2 – рентабельность производства;

X_3 – фондоотдача;

X_4 – общие затраты на 1 у.е. произведенной продукции;

X_5 – удельный вес рабочих на предприятии.

Снова предполагаем нормальность распределения всего пятимерного вектора признаков.

По результатам первичного статистического анализа были найдены следующие выборочные моменты:

		X_1	X_2	X_3	X_4	X_5
		Средние				
$x_{j\bullet} =$		50.70	30.36	137.64	55.01	60.24
		Стандартные отклонения				
$s_j =$		5.52	3.64	7.64	3.74	5.12
		Матрица ковариаций (l_{jk})				
	X_1	30.434	3.864	22.407	4.388	18.698
	X_2	3.864	13.273	7.090	-2.363	-0.618
$L =$	X_3	22.407	7.090	58.327	22.819	13.842
	X_4	4.388	-2.363	22.819	14.025	8.608
	X_5	18.698	-0.618	13.842	8.608	26.253

Обратная матрица ковариаций:

$$L^{-1} = \left(\tilde{l}_{jk} \right) = \begin{pmatrix} 0.288 & 0.159 & -0.291 & 0.550 & -0.228 \\ 0.159 & 0.227 & -0.233 & 0.448 & -0.132 \\ -0.291 & -0.233 & 0.379 & -0.708 & 0.234 \\ 0.550 & 0.448 & -0.708 & 1.416 & -0.473 \\ -0.228 & -0.132 & 0.234 & -0.473 & 0.229 \end{pmatrix}$$

Корреляционные связи.

Поскольку матрица корреляций симметрична, то для записи всех коэффициентов корреляции можно использовать только её верхнюю наддиагональную часть. В нижней половине матрицы можно поместить значения частных корреляций r_{kj*} , вычисляемые по обратной матрице ковариаций в соответствии с формулой [10, с.14, замечание 9]:

	матрица корреляций (частных корреляций)				
	X_1	X_2	X_3	X_4	X_5
X_1		0.19	0.53	0.21	0.66
X_2	-0.62		0.25	-0.17	-0.03
X_3	0.88	0.79		0.80	0.35
X_4	-0.86	-0.79	0.97		0.45
X_5	0.89	0.58	-0.80	0.83	

Прежде чем делать какие-либо выводы, необходимо проверить все эти корреляции на значимость с помощью преобразования Стьюдента. В

следующей таблице приведены p -значения при двусторонней альтернативе для полных и частных корреляций.

	p -значения для корреляций				
	X_1	X_2	X_3	X_4	X_5
X_1		0.18	<0.001	0.14	<0.001
X_2	<0.001		0.07	0.23	0.82
X_3	<0.001	<0.001		<0.001	0.012
X_4	<0.001	<0.001	<0.001		0.001
X_5	<0.001	<0.001	<0.001	<0.001	

Анализируя таблицу корреляций можно обнаружить как ложные, так и скрытые зависимости.

СКРЫТАЯ ЗАВИСИМОСТЬ			
Признаки	Полная корреляция		Частная корреляция
$X_1 - X_2$:	незначимая >0	$\mapsto$	высоко значимая <0
$X_1 - X_4$:	незначимая >0	$\mapsto$	высоко значимая <0
$X_2 - X_3$:	слабо значимая >0	$\mapsto$	высоко значимая >0
$X_2 - X_4$:	незначимая <0	$\mapsto$	высоко значимая <0
$X_2 - X_5$:	незначимая $=0$	$\mapsto$	высоко значимая >0
ЛОЖНАЯ ЗАВИСИМОСТЬ			
$X_3 - X_5$:	значимая >0	$\mapsto$	высоко значимая <0

Другими словами, если ориентироваться на полный коэффициент корреляции $r_{35} = 0.35$, можно сделать опрометчивый вывод, что с увеличением признака X_3 следует ожидать увеличения признака X_5 . В действительности это было бы так, если бы при этом все остальные признаки изменялись случайно. Если же их значения зафиксированы, то как раз наоборот, увеличивая X_3 , мы, скорее всего, добьемся уменьшения X_5 . С другой стороны, при кажущейся независимости X_2 и X_5 ($r_{25} \approx 0$), можно ожидать увеличения признака X_2 , если увеличивать значение признака X_5 , оставляя неизменными значения остальных признаков.

Для некоторых из коэффициентов корреляции построим доверительные интервалы Фишера (I.4, стр.17):

	Доверительные интервалы		
	для ρ_{15} (при $r_{15} = 0.66$)	для ρ_{23} (при $r_{23} = 0.25$)	для ρ_{34*} (при $r_{34*} = 0.97$)
Фишера	(0.472 ; 0.793)	(−0.022 ; 0.496)	(0.941 ; 0.981)
Точные	(0.466 ; 0.790)	(−0.025 ; 0.494)	(0.941 ; 0.981)

В последней строке таблицы для сравнения приведены значения точных доверительных интервалов, вычисленные с использованием точного распределения выборочного коэффициента корреляции (см. теорему I.15, стр.30). Как видно, расхождения весьма незначительные.

Второй из построенных интервалов подтверждает вывод о слабой значимости коэффициента $r_{23} = 0.25$. Заметим, кроме того, что при вычислении доверительных интервалов для полных (Пирсоновских) коэффициентов корреляции мы полагали асимптотическую дисперсию равной $1/47$, а для частных коэффициентов корреляции равной $1/44$.

Наконец, проверим гипотезу независимости сразу всех признаков по критерию отношения правдоподобия (см. следствие 3, стр.38). Статистика критерия (логарифм определителя матрицы корреляций), число степеней свободы и вспомогательный параметр равны

$$-\ln(|R|) = 0.018, \quad \nu = \frac{5 \cdot 4}{2} = 10, \quad \eta = 50 - \frac{2 \cdot 6 + 11}{6} = 46.5.$$

Таким образом, критический уровень значимости

$$\alpha_{кр} = 1 - \mathcal{K}_{10}(46.5 \cdot 4.02) < 10^{-30}.$$

Гипотеза независимости всех признаков должна быть отвергнута.

Множественная корреляция.

Прежде чем переходить к построению функций линейной регрессии, вычислим коэффициенты множественной корреляции и детерминации. Для этого снова воспользуемся обратной матрицей ковариаций. Например,

$$r_{1*}^2 = 1 - \frac{1}{l_{11}\tilde{l}_{11}} = 1 - \frac{1}{30.434 \cdot 0.288} = 0.886 = 0.94^2.$$

Остаточная дисперсия, то есть дисперсия ошибки линейного прогноза X_1 по всем остальным признакам, равна

$$s_{1*}^2 = 1/\tilde{l}_{11} = 1/0.288 = 3.474.$$

Коэффициенты для других признаков вычисляются аналогично.

Множественные корреляции					
r_{j*}	0.94	0.82	0.98	0.97	0.91
Коэффициенты детерминации					
$r_{j*}^2 \cdot 100\%$	89	67	95	95	83
Остаточные дисперсии					
s_{j*}^2	3.474	4.409	2.642	0.706	4.367

Таким образом, например, рентабельность производства на 67% обусловлена влиянием на неё остальных четырех рассматриваемых характеристик.

При таких высоченных значениях коэффициентов мало смысла проверять их на значимость отличия от нуля. Мы проделаем это для самого маленького из них $r_{2*} = 0,82$ исключительно для полноты картины. Статистика Фишера для этого коэффициента равна (I.21, стр.39)

$$F = \frac{45}{4} \cdot \frac{0.82^2}{1 - 0.82^2} = 23.09.$$

По распределению Фишера с параметрами $(4, 45)$ находим критический уровень значимости

$$\alpha_{\text{кр}} = 1 - \mathcal{F}_{4,45}(23.09) = 0.000\,000\,000\,2,$$

что значительно меньше любого мыслимого уровня значимости.

Прогноз рентабельности производства по всем признакам.

Построим теперь линейный прогноз X_2 по значениям остальных признаков (X_1, X_3, X_4, X_5) . Коэффициенты этого прогноза равны

Коэффициенты регрессии b_{2j*}					
	b_{21*}	C	b_{23*}	b_{24*}	b_{25*}
b_{2j*}	-0.701	-1.68	1.025	-1.975	0.583
Стандартные ошибки коэффициентов					
$\frac{\hat{\tau}_\beta}{\sqrt{n-p}}$	0.131	6.6	0.117	0.228	0.122

Опять же эти коэффициенты вычислялись через элементы $\tilde{l}_{jk}$ обратной матрицы ковариаций L^{-1} (см. [10, стр.7]). Например, коэффициент

$$b_{21*} = -\frac{\tilde{l}_{21}}{\tilde{l}_{22}} = -\frac{0.159}{0.127} = -0.701.$$

На месте коэффициента b_{22*} здесь проставлен свободный член регрессии C , который равен произведению, взятому с обратным знаком, строки коэффициентов регрессии $(b_{21*}, -1, b_{23*}, b_{24*}, b_{25*})$ на столбец всех выборочных средних:

$$C = b_{21*} x_{1\cdot} + (-1)x_{2\cdot} + b_{23*} x_{3\cdot} + b_{24*} x_{4\cdot} + b_{25*} x_{5\cdot}.$$

Для вычисления стандартных ошибок коэффициентов нужно построить информационную матрицу $\mathbb{S}$ и обратную к ней $\mathbb{S}^{-1}$. Этот процесс хоть и не сложен, но занимает много места, поэтому здесь мы его опустим. Отметим только, что дисперсии всех коэффициентов регрессии, вплоть до свободного члена, связаны с соответствующими диагональными элементами $\tilde{s}_{jj}$ матрицы $\mathbb{S}^{-1}$ соотношением

$$\tau_{\beta_j}^2 = n s_{2*}^2 \tilde{s}_{jj}.$$

При $j = 2$ (номер параметра, рассматриваемого как отклик) эта формула дает дисперсию остаточного члена регрессии.

В нашем примере диагональ матрицы $\mathbb{S}^{-1}$ равна
(0.00353, 8.890, 0.00280, 0.0106, 0.00304).

Поэтому, например,

$$\frac{\tau_{\beta_1}}{\sqrt{n-p}} = \sqrt{\frac{50 \cdot 4.409 \cdot 0.00353}{45}} = 0.131.$$

Значимость всех коэффициентов регрессии вытекает из значимости частных коэффициентов корреляции (см. [10, стр.113, замечание 10]). Другой способ проверки равенства нулю коэффициента регрессии состоит в построении доверительного интервала для этого коэффициента. Например, верхняя 2.5%-ая точка нормального закона равна 1.96, поэтому 95%-доверительный интервал для коэффициента β_{21*} равен

$$b_{21*} \mp 1.96 \frac{\tau_{\beta_1}}{\sqrt{n-p}} = -0.701 \mp 1.96 \cdot 0.131 = (-0.96; -0.44).$$

Так как этот интервал не содержит точку 0, то на уровне $\alpha = 0.05$ гипотеза о равенстве нулю коэффициента регрессии β_{21*} должна быть отвергнута.

Доверительный интервал для свободного члена регрессии C будет содержать нулевую точку при любом доверительном уровне, большем 20%. Другими словами, минимальный уровень значимости, при котором гипоте-

за о равенстве нулю истинного значения C будет отвергнута, равен 0.8. Таким образом, скорее всего истинное значение $C = 0$.

З 1. Несмотря на то, что свободный член регрессии незначим, его нельзя просто так выбросить из уравнения регрессии, поскольку построение наилучшего прогноза с помощью линейной функции, выходящей из нуля, это другая задача и для неё получатся другие формулы коэффициентов. Кстати, как-нибудь на досуге решите эту задачу **самостоятельно**.

Теперь найдем прогноз признака X_2 по конкретным значениям остальных признаков. В качестве примера возьмем строку реальных данных

$$x_1 = 50.2, x_2 = 25.7, x_3 = 135.1, x_4 = 57, x_5 = 63.1$$

и сравним действительное значение $x_2 = 25.7$ с наилучшим линейным прогнозом. Для удобства вычислений заменим значение x_2 в этой строке на единицу и умножим вектор $\vec{X}_0 = (50.2, 1, 135.1, 57, 63.1)'$ скалярно слева на столбец коэффициентов регрессии (вместе с C). В результате получим прогноз

$$\hat{x}_{2*} = 25.82,$$

который мало отличается от действительно имевшего место значения x_2 .

Дисперсия оценки наилучшего прогноза (I.13, стр.27) равна

$$\frac{1}{n-p} \tau_{\eta}^2 = 0.35.$$

Таким образом, 95%-доверительный интервал для $\eta(\vec{X}_0)$

$$(25.82 \mp 1.96 \cdot \sqrt{0.35}) = (24.66; 26.98).$$

При построении толерантных пределов возможных значений X_2 (см. замечание З16, стр.21) необходимо дисперсию $\frac{1}{n-p} \tau_{\eta}^2$ заменить на

$$\frac{1}{n-p} \tau_{\eta}^2 + s_{2*}^2 = 0.35 + 4.409 = 4.759.$$

Таким образом, при значениях $(X_1, X_3, X_4, X_5) = (50.2, 135.1, 57, 63.1)$ толерантный 95%-интервал значений рентабельности производства равен

$$(20.38; 31.25).$$

Если прогноз значений X_2 планируется осуществлять во всех возможных точках 4-мерного вектора предикторов (X_1, X_3, X_4, X_5) , то при

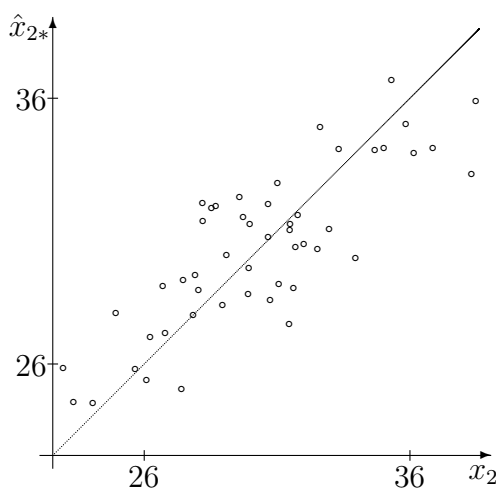
построении доверительного интервала для $\eta(\vec{X}_0)$ константу 1.96 необходимо заменить на

$$C_n^{(\alpha)} = \sqrt{5F_{5,45}^{(0.05)}} = 3.48.$$

В результате получим доверительный интервал Уоркинга-Хотеллинга

$$(25.82 \mp 3.48 \cdot \sqrt{0.35}) = (24.76; 27.88).$$

Качество линейного прогноза можно графически проиллюстрировать, если для всех выборочных данных сравнить прогноз признака $\hat{x}_{2*}$ с действительно имевшими место значениями x_2 . Кроме всех пар точек $(x_2, \hat{x}_{2*})$ на график ещё наносится прямая линия $\hat{x}_{2*} = x_2$, по которой можно судить о зависимости ошибки прогноза X_2 от величины истинного значения X_2 .



Поскольку на приведенном здесь графике разброс точек по оси ординат одинаков на всем протяжении прямой линии, то можно сделать вывод о независимости остаточной дисперсии от значений признака X_2 , что, в свою очередь, служит хорошим подтверждением линейности регрессии.

Прогноз рентабельности производства по общим затратам.

Построим наилучший прогноз значения признака X_2 только по одному из признаков X_4 , с которым X_2 имеет незначимую корреляцию $r_{24} = -0.17$. Коэффициенты этого прогноза равны (см.[10,с.8, (I.2)])

Коэффициенты регрессии		Стандартные ошибки		Коэффициент корреляции - детерминации	
C	39.624	$\frac{\tau_C}{\sqrt{n-2}}$	7.625	r	-0.17
b_{24}	-0.168	$\frac{\tau_b}{\sqrt{n-2}}$	0.138	$r^2 \cdot 100\%$	3.0%

Стандартная ошибка для коэффициента b_{24} вычислена по формуле (I.7, стр.18) с уточнением, приведенным на стр.24. Для отыскания стандартной ошибки S использована формула (I.8, стр.19), в которой значение предикторной переменной положено равным нулю.

§ 2. Границы 95%-доверительного интервала для коэффициента регрессии β_{24} равны $-0.168 \pm 1,96 \cdot 0.138$. Таким образом, доверительный интервал для β_{24} — $(-0.44; 0.103)$ — накрывает точку 0, что говорит о его незначимости. Опять же, отсюда не следует, что наилучший прогноз признака X_2 по значению признака X_4 равен всегда константе $C = 39.624$. Построение наилучшего прогноза сл.в. посредством константы — это другая задача, и её решение хорошо известно — выборочное среднее:

$$\hat{x}_2 = x_{2\bullet} = 30,36.$$

Стандартная ошибка такого прогноза равна $\frac{s_2}{\sqrt{n-1}} = 0.52$.

§ 3. Подбором легко находится значение $\alpha = 0.23$, при котором доверительный интервал для β_{24} не будет накрывать точку 0. Следовательно, критический уровень значимости при проверке гипотезы **H** о равенстве нулю коэффициента β_{24} равен 0.23, что идеально совпадает с критическим уровнем значимости $\alpha_{кр} = 0.23$ при проверке эквивалентной гипотезы о равенстве нулю коэффициента корреляции.

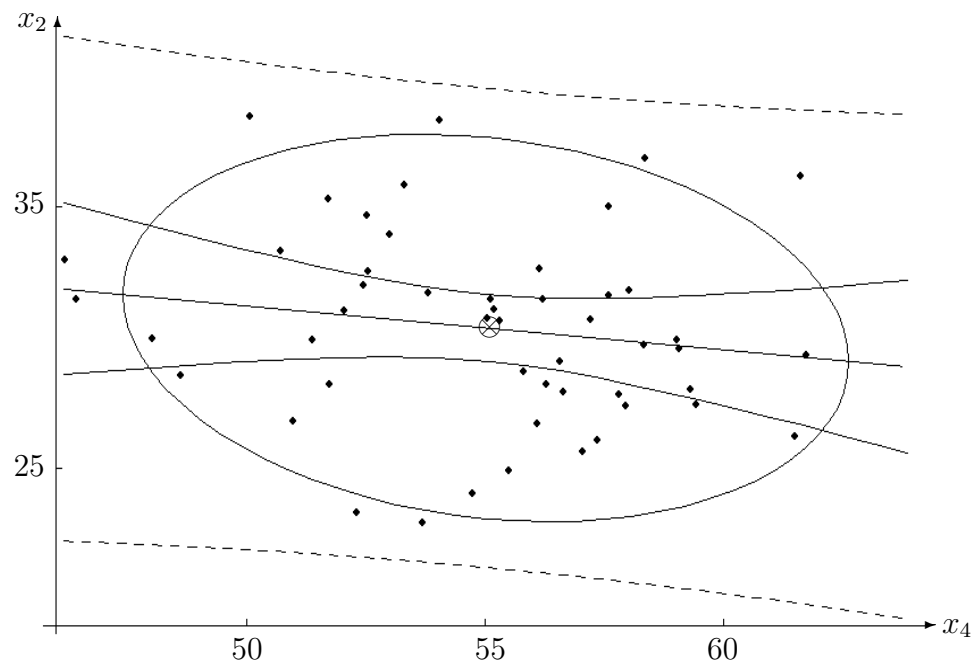
На приведенном ниже рисунке нанесены

- а) линии регрессии [10, стр.8, (I.2)];
- б) 95%-доверительные и толерантные границы для регрессии (I.12, стр.19);
- в) эллипс рассеяния [10, стр.23, (I.6)];
- г) все пары точек исходных данных признаков X_2, X_4 ;
- д) центральная точка выборочных данных.

Широченный разброс точек вокруг линии прогноза, как и коэффициент детерминации 3%, показывают, что мало проку в использовании признака X_4 в качестве прогностического фактора для признака X_2 .

§ 4. Двумерный нормальный сл.вектор с вероятностью 0,135 не попада-

ет внутрь своего эллипса рассеяния [10, стр.23]. Среди наших данных эти точки составили 16%, что можно считать неким подтверждением нормальности распределения выборки.



Каноническая корреляция.

Первые два признака наших данных можно рассматривать как целевые факторы производства. Представляет интерес исследование влияния на оба эти признака всех остальных признаков.

Прежде всего необходимо решить вопрос о принципиальной возможности такого влияния. Другими словами, сначала надо проверить гипотезу независимости векторов $\vec{U} = (X_1, X_2)'$ и $\vec{V} = (X_3, X_4, X_5)'$. Для этого снова воспользуемся критерием хи-квадрат (I.19, стр.37). Оценки обобщенных дисперсий рассматриваемых векторов равны (см. ниже разбиение матрицы корреляций)

$$\hat{\mathfrak{S}}^2(\vec{U}) = |S_{11}| = 389.00, \quad \hat{\mathfrak{S}}^2(\vec{V}) = |S_{22}| = 6234.47.$$

Статистика критерия отношения правдоподобия

$$\hat{\mathfrak{Z}} = \ln(389.00) + \ln(6234.47) - \ln(155834.68) = 2.75$$

с числом степеней свободы $\nu = \frac{(5^2 - (2^2 + 3^2))}{2} = 6$

и параметром $\eta = 50 - 1, 5 - \frac{5^4 - 5 \cdot 6}{48} - \frac{5^3 - (2^3 + 3^3)}{6 \cdot 6} = 46.$

Таким образом, критический уровень значимости равен

$$\alpha_{\text{кр}} = 1 - K_6(46 \cdot 2, 75) < 10^{-10}.$$

Следовательно, гипотеза независимости первых двух признаков от последних трех признаков должна быть отвергнута.

В первой части курса [10, стр.19, Теорема I.6] приведен способ построения канонических переменных и вычисления канонических корреляций. Чтобы нивелировать возможное влияние единиц измерения признаков, мы применим этот способ не к матрице ковариаций, а к матрице корреляций. При этом изменятся только коэффициенты канонических переменных, но не сами канонические корреляции. Необходимо найти собственные числа двумерной матрицы

$$K = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21} = \begin{pmatrix} 0.767 & 0.314 \\ 0.243 & 0.398 \end{pmatrix},$$

где матрицы $R_{11}, R_{12}, R_{21}, R_{22}$ выбраны из полной матрицы корреляций по следующей схеме:

$$\left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array} \right) = \left(\begin{array}{cc|ccc} 1 & 0.19 & 0.53 & 0.21 & 0.66 \\ 0.19 & 1 & 0.25 & -0.17 & -0.03 \\ \hline 0.53 & 0.25 & 1 & 0.80 & 0.35 \\ 0.21 & -0.17 & 0.80 & 1 & 0.45 \\ 0.66 & -0.03 & 0.35 & 0.45 & 1 \end{array} \right).$$

Корни характеристического уравнения для матрицы K

$$(0.767 - q)(0.398 - q) = 0.243 \cdot 0.314$$

равны квадратам канонических корреляций:

$$r_1^2 = 0.914, \quad r_2^2 = 0.250.$$

Для отыскания собственного вектора, отвечающего собственному числу $q = r_1^2 = 0.914$, необходимо найти нормированное решение уравнения

$$(0.767 - 0.914)a_1 + 0.314a_2 = 0.$$

Полагая сначала $a_1 = 1$, находим $a_2 = 0.469$. Теперь вычислим дисперсию линейного преобразования $a_1 X_1 + a_2 X_2$:

$$d^2 = (a_1, a_2) R_{11} (a_1, a_2)' = 1.401$$

Нормированный (на $d = 1.184$) собственный вектор равен

$$(a_1, a_2) = \left(\frac{1}{d}, \frac{0.469}{d} \right) = (0.845, 0.397).$$

Коэффициенты линейного преобразования вектора $(X_3, X_4, X_5)'$ вычисляются как

$$\frac{1}{r_1} R_{22}^{-1} R_{21} (a_1, a_2)' = (1.337, -1.245, 0.657).$$

Таким образом, первая каноническая корреляция и соответствующие ей линейные преобразования исследуемых векторов равны

$$r_1 = 0.956, \quad \begin{cases} U_1 = 0.845 X_1 + 0.397 X_2, \\ V_1 = 1.337 X_3 - 1.245 X_4 + 0.657 X_5. \end{cases}$$

В компоненте U_1 основная нагрузка приходится на признак X_1 – заработная плата, а во второй на два признака X_3 – фондоотдача и X_4 – затраты на единицу произведенной продукции (с противоположным знаком). Если этим признакам придать какое-либо броское наименование, например, „экономичность производства“, то можно было бы сказать, что доходы работников предприятий напрямую зависят от экономичности производства.

Для второй канонической корреляции имеем

$$r_2 = 0.50, \quad \begin{cases} U_2 = -0.570 X_1 + 0.939 X_2, \\ V_2 = 0.463 X_3 - 0.502 X_4 - 0.365 X_5. \end{cases}$$

Поскольку все признаки в компоненте V_2 несут почти одинаковую нагрузку, то вторая компонента представляет собой интегральную характеристику условий производства – „отлаженности“. Следовательно, рентабельность производства умеренно коррелирует с отлаженностью производства.

Главные компоненты.

Для построения главных компонент необходимо найти собственные числа всей пятимерной матрицы ковариаций (см. [10, стр.45, Теорема II.1]). Собственные числа и матрица преобразований (собственные векторы) равны

$$\begin{array}{cc|ccc} \gamma & \% & C' & & & & \\ \hline 87.48 & 62\% & 0.44 & 0.08 & 0.76 & 0.30 & 0.35 \\ 28.55 & 82\% & 0.55 & -0.08 & 0.50 & -0.24 & 0.62 \\ 17.70 & 94\% & 0.36 & 0.72 & 0.07 & -0.47 & -0.36 \\ 8.16 & 99\% & 0.52 & -0.64 & 0.08 & -0.15 & -0.54 \\ 0.43 & 100\% & 0.32 & 0.25 & -0.40 & 0.78 & -0.27 \end{array}.$$

В соответствии с принятой нами терминологией основную информацию о разбросе данных – 82%, несут первые две главные компоненты,

причем в первой основная нагрузка приходится на показатель X_3 – фондоотдача, а во второй на три показателя X_1, X_3, X_5 . Снова, как и при построении канонических корреляций, этой группе признаков следовало бы придумать некое звучное имя, которое впоследствии использовать, например, для сравнения предприятий в различных отраслях.

Вычислив для всех данных первые две главные компоненты, можно представить их графически, что позволяет иногда разделить всю совокупность данных на несколько групп. В рассматриваемом примере точки на графике располагаются достаточно *т*кучно, поэтому мы не приводим здесь этот график.

Пример IV. *Дискриминантный анализ больных гипертериозом.*

Изучалась возможность прогноза результатов лечения больных гипертериозом (увеличение щитовидной железы) по результатам теста на радиоактивный йод. На каждом пациенте проводилось десять таких тестов (пять до лечения – признаки X_1, X_2, X_3, X_4, X_5 , и пять во время лечения – признаки $X_6, X_7, X_8, X_9, X_{10}$). Всего было обследовано 23 пациента. После лечения пациенты были разделены на три группы.

ГРУППА 1. Лечение оказалось успешным; проведенное через большой промежуток времени клиническое обследование показало, что пациент здоров (16 пациентов).

ГРУППА 2. Лечение безуспешно, то есть состояние больного осталось без изменения (4 пациента).

ГРУППА 3. Исход лечения успешен, но в дальнейшем возможен рецидив (3 пациента).

Требуется проверить предположение о том, что исход лечения зависит от значений десяти указанных измерений (то есть может быть предсказан по ним). Кроме того, необходимо предложить способ отнесения любого пациента к той или иной группе.

В рассматриваемом эксперименте

число групп $J = 3$;

число признаков $p = 10$;

объемы выборок в группах $n_1 = 16, n_2 = 4, n_3 = 3$;

общий объем наблюдений $n = 23$.

векторы средних значений

		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
группа 1	$\vec{x}_1.$	89.3	90.6	83.8	70.7	19.0	31.1	37.2	43.9	411.0	2.46
группа 2	$\vec{x}_2.$	76.4	78.2	76.9	69.3	18.3	68.3	69.1	60.0	54.5	15.5
группа 3	$\vec{x}_3.$	85.3	87.7	73.6	68.2	30.1	29.2	38.5	50.1	48.2	1.07
общий	$\vec{x}_{..}$	86.5	88.1	81.3	70.2	20.3	37.4	42.9	47.5	44.3	4.55

матрица ковариаций L (несмещенная оценка)

$$\begin{pmatrix} 239.8 & 163.8 & 142.7 & 138.6 & 15.7 & 114.0 & 115.6 & 116.5 & 93.3 & 0.88 \\ 163.8 & 191.9 & 169.4 & 170.3 & -12.3 & 85.5 & 83.3 & 70.4 & 59.7 & -0.57 \\ 142.7 & 169.4 & 192.1 & 207.2 & -37.9 & 79.9 & 90.1 & 91.1 & 78.4 & -2.4 \\ 138.6 & 170.3 & 207.2 & 307.2 & -58.7 & 79.6 & 82.6 & 96.5 & 77.6 & 1.40 \\ 15.7 & -12.3 & -37.9 & -58.7 & 104.3 & 28.2 & 44.1 & 39.3 & 43.3 & 4.01 \\ 114.0 & 85.5 & 79.9 & 79.6 & 28.2 & 211.8 & 221.3 & 245.2 & 226.6 & -1.08 \\ 115.6 & 83.3 & 90.1 & 82.6 & 44.1 & 221.3 & 261.4 & 280.6 & 268.9 & -2.52 \\ 116.5 & 70.4 & 91.1 & 96.5 & 39.3 & 245.2 & 280.6 & 371.8 & 332.4 & -2.28 \\ 93.3 & 59.7 & 78.4 & 77.6 & 43.3 & 226.6 & 268.9 & 332.4 & 331.6 & 0.98 \\ 0.88 & -0.57 & -2.4 & 1.40 & 4.01 & -1.08 & -2.52 & -2.28 & 0.98 & 12.2 \end{pmatrix}$$

Сравнение всех групп.

Статистика Фишера (III.5, стр.73) и степени свободы при проверке гипотезы

$H : \vec{\mu}_1 = \vec{\mu}_2 = \vec{\mu}_3$ о совпадении центров всех групп равны

$$T^2 = 5.3, \quad g_1 = 200, \quad g_2 = 11.$$

Таким образом, критический уровень значимости

$$\alpha_{кр} = 1 - \mathcal{F}_{200,11}(5.3) = 0.002.$$

Вывод. Гипотеза неразличимости всех трех групп должна быть отвергнута.

З 5. Применяя критерий Λ -Уилкса (З6, стр.74), получим такое же значение критического уровня значимости. В рассматриваемом примере только два собственных числа отличны от нуля

$$\lambda_1 = 9.08, \quad \lambda_2 = 0.55.$$

Следовательно, статистика

$$\Lambda = \frac{1}{10.08} \cdot \frac{1}{1.55} = 0.064,$$

$$-\left(n - \frac{p+J+2}{2}\right) \ln \Lambda = 42.58, \quad \alpha_{кр} = 1 - \mathcal{K}_{20}(42.58) = 0.002.$$

Попарное сравнение групп.

Визуально можно заметить, что различие обусловлено, в основном, второй группой пациентов. Для проверки этого предположения сравним группы попарно методом Шеффе (III.7, стр.75):

Группы	Статистика T^2	p	Различие
1 – 2	8.82	0.004	значимо
1 – 3	0.91	0.98	незначимо
2 – 3	7.12	0.011	значимо
{1, 3} – 2	9.68	0.003	значимо

В последней строчке таблицы по той же схеме проведено сравнение второй группы с объединенными в одну группами 1 и 3.

Вывод. Тест на радиоактивный йод отделяет только тех больных, для которых лечение будет безуспешным.

Правила дискриминации.

Линейные функции, с помощью которых производится классификация индивидов по группам, основаны на собственных векторах, соответствующих собственным числам, полученным при построении критерия Λ -Уилкса в предыдущем замечании. Разделяющая способность этих функций оценивается посредством коэффициентов канонической корреляции:

$$R_1^* = \sqrt{\frac{9.08}{1 + 9.08}} = 0.95 \quad R_2^* = \sqrt{\frac{0.55}{1 + 0.55}} = 0.59.$$

Таким образом, первая дискриминантная функция обладает высочайшей разделительной способностью.

Сами дискриминантные функции задаются векторами линейных преобразований:

$$\vec{e}'_1 = (-0.021, -0.078, 0.072, -0.011, -0.0050, 0.109, 0.037, -0.021, -0.074, 0.1817),$$

$$\vec{e}'_2 = (-0.027, 0.107, -0.192, 0.074, 0.019, -0.065, 0.072, 0.023, -0.0089, 0.023).$$

Процесс отнесения индивида в ту или иную группу продемонстрируем на конкретном пациенте, попавшем в группу 1:

$$\vec{x}' = (56.7, 54.6, 49.2, 40.5, 22.0, 20.1, 28.2, 39.0, 40.1, 1.1).$$

1) Для каждой из групп находим средние значения (центры) всех дискриминантных функций и квадраты радиусов шаров вокруг этих центров:

$$\vec{w}'_{1\bullet} = (-2.48, -1.89), \quad \vec{w}'_{2\bullet} = (4.57, -1.24), \quad \vec{w}'_{3\bullet} = (-4.01, 0.16),$$

$$r_1^2 = 3.52 \cdot \frac{t(n_1 + 1)(n - J)}{(n - J - t + 1)n_1} = 7.88, \quad r_2^2 = 9.27, \quad r_3^2 = 9.89.$$

Здесь число 3.52 предстает собой верхнюю 5%-ую квантиль распределения Фишера с параметрами (2,19), а количество ненулевых собственных чисел $t = 2$.

2) Находим значения дискриминантных функций для рассматриваемого индивида: $\omega_1 = \vec{e}'_1 \vec{x} = -2.78$, $\omega_2 = \vec{e}'_2 \vec{x} = -0.46$.

3) Вычисляем квадраты расстояний вектора $\vec{\omega} = (\omega_1, \omega_2)$ до центров дискриминантных функций всех групп:

$$v_1 = \|\vec{\omega} - \vec{w}_{1\bullet}\|^2 = 2.13, \quad v_2 = \|\vec{\omega} - \vec{w}_{2\bullet}\|^2 = 54.68, \quad v_3 = \|\vec{\omega} - \vec{w}_{3\bullet}\|^2 = 1.89.$$

4) Сравнивая полученные расстояния с радиусами, делаем вывод, что рассматриваемый индивид может быть отнесен как к первой, так и к третьей группе. На этом можно было бы остановиться, поскольку ранее нами был сделан вывод о неразличимости групп 1 и 3. Однако далее в иллюстративных целях мы попытаемся уточнить наш вывод. Здесь же ещё отметим, что такой способ выделяет в пространстве дискриминантных функций области классификации в виде шаров вокруг центров групп. На следующем рисунке эти шары (круги на плоскости при трех группах) приведены вместе с дискриминантными функциями для всех 23 пациентов (символ " Δ " использован для пациентов группы 1, символ " $*$ " – для группы 2, символ " $\square$ " – для группы 3, символ " $\otimes$ " – для центров групп, символ " $\spadesuit$ " – для исследованного в примере пациента первой группы).

5) Однозначной классификации можно добиться, относя индивида

а) к ближайшей группе;

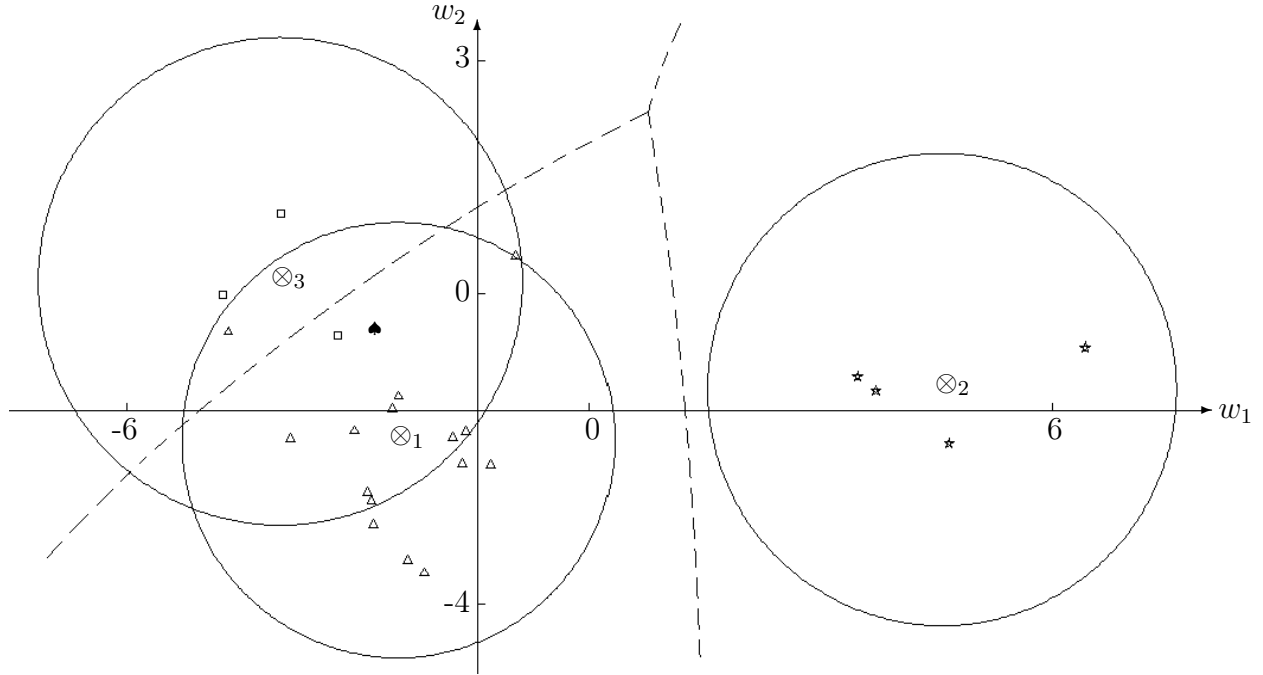
б) к группе с наименьшим значением статистики

$$q_j = \frac{(n - J - t + 1)n_j}{t(n - J)(n_j + 1)} v_j, \quad j = 1, 2, 3;$$

в) к группе с наименьшим значением статистики

$$l_j = \left(1 + \frac{n_j}{(n - J)(n_j + 1)} v_j\right) \frac{1}{\sqrt[n+1]{P_j^2}}, \quad j = 1, 2, 3,$$

где P_j – доля индивидов из j -ой группы (здесь как-бы учитывается распределение больных по группам в „генеральной совокупности“).



Для нашего случая при втором способе дискриминации

$$q_1 = 0.95, \quad q_2 = 20.78, \quad q_3 = 0.67,$$

то есть для этого пациента в дальнейшем следует ожидать возможного рецидива болезни. Аналогичный результат даёт первый способ.

При третьем способе $l_1 = 1.13$, $l_2 = 3.6$, $l_3 = 1.27$, следовательно, можно надеяться, что пациент окончательно выздоровел.

Все три способа классификации выделяет в пространстве дискриминантных признаков J областей. На нашем рисунке пунктирной линией представлены области, порождаемые третьим правилом дискриминации.

Из рисунка видно также, что только по одному пациенту из групп 1 и 3 были ошибочно квалифицированы. Умиляться по этому поводу особенно не стоит, поскольку, как показывает опыт, такой способ оценки точности правила дискриминации слишком оптимистичен.

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

1. Проверьте равенство

$$\mathbf{Cov}(((X_1 - \mu_1)^2, (X_2 - \mu_2)^2, (X_1 - \mu_1)(X_2 - \mu_2))) = Q$$

для матрицы Q на стр.7.

2. Докажите, что если последовательность случайных векторов $\vec{X}_n \rightsquigarrow F(\vec{x})$, то для любой последовательности констант $c_n \rightarrow 1$

$$c_n \vec{X}_n \rightsquigarrow F(\vec{x}).$$

3. Докажите состоятельность оценки l_{jk} .

4. Докажите многомерный вариант дельта-метода.

5. Найдите и сформулируйте теорему Слуцкого.

6. Обоснуйте следующее утверждение:

если распределение статистики T приближенно совпадает с нормальным распределением со средним θ и дисперсией σ^2 , то интервал $T \pm 2 \cdot \sigma$ будет приближенным 95%-ым доверительным интервалом для параметра θ .

7. Проверьте соотношение

$$\operatorname{arcth}(x) = \frac{1}{2} \ln \left(\frac{1+x}{1-x} \right).$$

8. Постройте критерии проверки гипотез $\mathbf{H} : \rho = \rho_0$ и $\mathbf{H} : \rho \leq \rho_0$ об истинном коэффициенте корреляции по двумерной выборке

- а) из нормальной совокупности;
- б) из не гауссовской совокупности.

9. Постройте односторонние доверительные интервалы для истинного коэффициента корреляции по двумерной выборке

- а) из нормальной совокупности;
- б) из не гауссовской совокупности.

10. Докажите теорему I.17, стр.33
11. Докажите формулу (I.14, стр.27), связывающую информационную матрицу и матрицу ковариаций.
12. Проверьте равенство
- $$\frac{1}{2}(p^2 - \sum p_j^2) = \left[p + \frac{1}{2}p(p+1) \right] - \left[p + \sum \frac{1}{2}p_j(p_j+1) \right],$$
- заявленное в замечании Z23, стр.37.
13. Докажите соотношение между коэффициентом конкордации Кендэлла и коэффициентом корреляции Спирмена (замечание Z2, стр.63).
14. Проверьте (по индукции) равенство (II.3, стр.65).
15. В предположении независимости признаков опишите точное распределение коэффициента конкордации W , построенного по четырем наблюдениям с тремя признаками.
16. Найдите выражения для коэффициентов регрессии в ситуации, когда известно, что свободный член регрессии равен нулю.

СПИСОК ОБОЗНАЧЕНИЙ

A'	–	транспонированная матрица A
$\dot{h}, \ddot{h}$	–	первая и вторая производные функции h
$\vec{X} = (X_1, \dots, X_p)'$	–	вектор p наблюдаемых признаков
$\mu_j = \mathbf{E} X_j$	–	истинное среднее признака X_j
$\lambda_{jk} = \mathbf{Cov}(X_j, X_k) = \mathbf{E}(X_j X_k) - \mu_j \mu_k$	–	истинная ковариация признаков X_j, X_k
$\Lambda = \mathbf{Cov}(\vec{X}) = \left(\lambda_{jk} \right)_{j,k=1}^p$	–	матрица ковариаций p -мерного вектора наблюдаемых признаков $\vec{X}$
$\sigma_j^2 = \mathbf{D} X_j = \lambda_{jj}$	–	истинная дисперсия признака X_j
$\rho_{jk} = \frac{\lambda_{jk}}{\sigma_j \sigma_k}$	–	истинный коэффициент корреляции признаков X_j, X_k
$x_{j\bullet} = \frac{1}{n} \sum_{i=1}^n x_{ji}$	–	выборочное среднее признака X_j
$l_{jk} = \frac{1}{n} \sum_{i=1}^n x_{ji} x_{ki} - x_{j\bullet} x_{k\bullet}$	–	выборочная ковариация признаков X_j, X_k
$L = \left(l_{jk} \right)_{j,k=1}^p$	–	матрица выборочных ковариаций вектора наблюдаемых признаков $\vec{X}$
$s_j^2 = l_{jj}$	–	выборочная дисперсия признака X_j
$r_{jk} = \frac{l_{jk}}{s_j s_k}$	–	выборочная корреляция признаков X_j, X_k (коэффициент корреляции Пирсона)
$\vec{Z} \sim \vec{Y} \quad (\mathcal{D})$	–	распределение вектора $\vec{Z}$ совпадает с распределением $\vec{Y}$ (описывается законом $\mathcal{D}$)
$\mathcal{D}$	–	каллиграфический шрифт для обозначения закона распределения

$\mathcal{N}_k(\vec{\mu} | \Lambda)$ — k -мерное нормальное распределение с вектором средних $\vec{\mu}$ и матрицей ковариаций Λ

$\mathcal{N}^{(\alpha)}$ — верхняя α -квантиль стандартного нормального распределения

$\mathcal{F}_{l,m}$ — распределение Фишера с (l, m) степенями свободы

$\mathcal{F}_{l,m}^{(\alpha)}$ — верхняя α -квантиль распределения Фишера с (l, m) степенями свободы

$\mathcal{S}_m$ — распределение Стюдента с m степенями свободы

$t_m^{(\alpha)}$ — верхняя α -квантиль распределения Стюдента с m степенями свободы

$\mathcal{K}_m$ — хи-квадрат распределение с m степенями свободы

$\mathcal{K}_m^{(\alpha)}$ — верхняя α -квантиль хи-квадрат распределения с m степенями свободы

$\vec{Y}_n \rightsquigarrow \mathcal{N}_k(\vec{\mu} | Q)$ — последовательность векторов $\vec{Y}_n$ асимптотически нормальна с вектором средних $\vec{\mu}$ и матрицей ковариаций Q

$\vec{Y}_n \rightsquigarrow \mathcal{N}_k(\vec{\mu}_n | \frac{1}{n} Q)$ — нормированная разность $\sqrt{n}(\vec{Y}_n - \vec{\mu}_n)$ асимптотически нормальна $\mathcal{N}_k(\vec{0} | Q)$

(II.25) — ссылка на соответствующую формулу данного учебника

[5, (I.17)] — ссылка на формулу книги [5] из списка литературы

Счастье – это когда тебя понимают.

Из κ/φ

Хочешь быть счастливым – вызубри
сначала алфавит.

ГРЕЧЕСКИЙ АЛФАВИТ

Грек αλφαβετ

α : альфа	β : бета	γ : гамма	δ : дельта
ε : эпсилон	ζ : дзета	η : эта	λ : лямбда
ϑ : тета	θ : тета	μ : мю	ν : ню
ξ : кси	π : пи	ρ : ро	σ : сигма
τ : тау	φ : фи	χ : хи	ψ : пси
Γ : Гамма	Δ : Дельта	Θ : Тета	Λ : Лямбда
Σ : Сигма	Φ : Фи	Ψ : Пси	Ω : Омега

ГОТИЧЕСКИЙ АЛФАВИТ

Gothik Alfabet

Ɑ : А	Ɱ : Б	Ɐ : В	Ɒ : Ж
Ⱳ : Д	ⱳ : эФ	ⱴ : эЛЬ	Ⱶ : эН
ⱷ : эР	ⱸ : эС	ⱹ : Т	ⱺ : Икс

ЛИТЕРАТУРА

- [1] Андерсон Т. Введение в многомерный статистический анализ. – М.: „Физматгиз“, 1963. – 500 с.
- [2] Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ. – М.: „Финансы и статистика“, 1981. – 230 с.
- [3] Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: „Финансы и статистика“, 1998. – 350 с.
- [4] Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. – М.: „Наука“, 1973. – 900 с.
- [5] Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: „Наука“, 1976. – 736 с.
- [6] Крамер Г. Математические методы статистики. – М.: „Мир“, 1975. – 648 с.
- [7] Окунь Я. Факторный анализ. – М.: „Статистика“, 1974. – 200 с.
- [8] Себер Дж. Линейный регрессионный анализ. – М.: „Мир“, 1980. – 456 с.
- [9] Симушкин С.В. Дисперсионный анализ. Ч.1, Ч.2. – Казань.: „Издательство КГУ“, 1998.
- [10] Симушкин С.В. Многомерный статистический анализ. – Казань.: „Издательство КГУ“, 2006. – 98 с.
- [11] Шеффе Г. Дисперсионный анализ. – М.: „Наука“, 1980. – 512 с.

Симушкин Сергей Владимирович

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ. Часть II

Учебное пособие

Подписано в печать . .200 г.

Форм. бум. 60x84 1/16. Гарнитура „Таймс“. Печать ризографическая.

Печ.л. 7,125. Т.100. З.

Лаборатория оперативной полиграфии Издательства КГУ

420045, Казань, ул.Кр.Позиция, 2а

Тел. 231-52-12